Statistics and Operations Research Transactions, vol. 44, n. 2 (2020)

Invited article

SORT 44 (2) July-December 2020, 223-264 DOI: 10.2436/20.8080.02.101

Independent increments in group sequential tests: a review

KyungMann Kim[1,*] and Anastasios A. Tsiatis[2]

Abstract

In order to apply group sequential methods for interim analysis for early stopping in clinical trials, the joint distribution of test statistics over time has to be known. Often the distribution is multivariate normal or asymptotically so, and an application of group sequential methods requires multivariate integration to determine the group sequential boundaries. However, if the increments between successive test statistics are independent, the multivariate integration reduces to a univariate integration involving simple recursion based on convolution. This allows application of standard group sequential methods. In this paper we review group sequential methods and the development that established independent increments in test statistics for the primary outcomes of longitudinal or failure time data.

MSC: 62-02, 62H10, 62L10, 62L15.

Keywords: Failure time data, interim analysis, longitudinal data, clinical trials, repeated significance tests, sequential methods.

1 Introduction

In most chronic disease clinical trials, the primary outcome of interest is either longitudinal data taken at successive follow-up visits with possibly missing data or failure time data, i.e. time to an event such as death with possible right censoring. Typically participants enter the study serially in a way known as staggered entry, and the final analysis is conducted either after a pre-specified number of follow-up visits for each participant for longitudinal data or after a pre-specified follow-up period or a pre-specified number of events of interest for failure time data.

For ethical as well as practical reasons, these clinical trials are often monitored sequentially over time during the course of the study, and if a sufficiently large treatment difference is observed at an interim analysis, they may be considered for early stopping to avoid unnecessary experimentation on human subjects. Such an approach is known

[1] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, USA. kyungmann.kim@wisc.edu

[2] Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA. tsiatis@ncsu.edu

Received: 15 november 2020
Accepted: 30 november 2020

as a sequential method. When clinical trials are monitored in this way using a sequential method, multiplicity from repeatedly applying statistical testing over time has to be accounted for to control the overall type I error probability at an acceptable significance level. In order to determine the sequential boundaries that preserve the operating characteristic of a statistical test applied repeatedly, the joint distribution of test statistics over time has to be known.

For clinical trials in which the primary outcome of interest is taken only once from each participant, the joint distribution of test statistics over time is simply a product of the distributions of test statistics at each interim analysis as each participant contributes data to the test statistics only once and the increments between successive test statistics are independent. However, for the primary outcome that is either longitudinal or failure time data, it is no longer the case as each participant possibly contributes outcome data to test statistics more than once over interim analyses.

Modern-day clinical trials since the mid 1990s or even earlier have been routinely monitored by data and safety monitoring boards or data monitoring committees to ensure the safety of participants and whether risks versus benefits are acceptable for continuing the study. This is accomplished using standard group sequential methods in interim analyses for possible early stopping if there is clear statistical signal of differences in efficacy of an investigational intervention as compared to a control intervention that may include a placebo or standard of care or if there is major concerns for safety of participants. This review article on independent increments in group sequential tests is an attempt to describe the development of statistical methods for interim analyses leading up to mid 1990s.

For longitudinal data, the joint distribution of test statistics over time has been investigated by many including Armitage, Stratton and Worthington (1985), Geary (1988), Wei, Su and Lachin (1990a), Lee and DeMets (1991, 1992), Reboussin, Lan and DeMets (1992), Su and Lachin (1992), Wu and Lan (1992), Gange and DeMets (1996), and Lee, Kim and Tsiatis (1996). Likewise, for failure time data, the joint distribution of test statistics over time has been investigated by many including Tsiatis (1981, 1982), Gail, DeMets and Slud (1982), Slud and Wei (1982), Sellke and Siegmund (1983), Slud (1994), Tsiatis, Rosner and Tritchler (1985), Gu and Lai (1991), Lin (1992), Gu and Ying (1995), and Tsiatis, Boucher and Kim (1995).

Often the joint distribution turns out to be multivariate normal or at least asymptotically so, and subsequently sequential methods require multivariate numerical integration. The MULNOR program by Schervish (1984) can be used to this end, but it involves very intensive numerical computation. Also the program can handle multivariate integrations of only up to seven dimensions, thus limiting the tests to be applied up to seven times only.

If the increments between successive test statistics are independent, however, the multivariate numerical integration reduces to univariate numerical integration involving simple recursion based on convolution of two independent variables as noted by Armitage, McPherson and Rowe (1969) and McPherson and Armitage (1971). This is ob-

viously the case when the outcomes are measured only once as noted earlier. Moreover, this allows the use of standard group sequential methods such as by Pocock (1977), O'Brien and Fleming (1979), and Lan and DeMets (1983) for design and analysis of group sequential clinical trials.

The joint distributions established by these authors dealt with specific test statistics under selected statistical models for longitudinal data and failure time data. Jennison and Turnbull (1990) and Scharfstein, Tsiatis and Robins (1997), however, provided generalized theory for independent increments in sequential test statistics. The former considered the joint distribution of test statistics for treatment effect in the presence of covariates in regression model setting, while the latter considered the joint distribution of semiparametric-efficient test statistics.

The rest of this paper is organized as follows. In Section 2, we first review the historical development of sequential methods including classical and the so-called group sequential methods specifically for application in clinical trials as a background. We then review repeated significance testing and univariate recursive numerical integration when increments between successive test statistics are independent in contrast to the multivariate numerical integration required for sequential test statistics with correlated increments. In Section 3, we review the historical development for the joint distribution of sequential test statistics and independent increments for group sequential tests of longitudinal data and failure time data. In Section 4, after introducing general notations and formulation of the problem, we review joint distributions of sequentially computed test statistics for general regression models of independent data and various parametric, semiparametric and nonparametric models for longitudinal data and failure time data. In Section 5, we briefly review how the error spending function and information fraction is used for design and analysis of group sequential clinical trials and demonstrate independent increments in sequential test statistics for longitudinal data and failure time data using real clinical trials data and simulated data. We close with concluding remarks and observations in Section 6.

2 Sequential methods

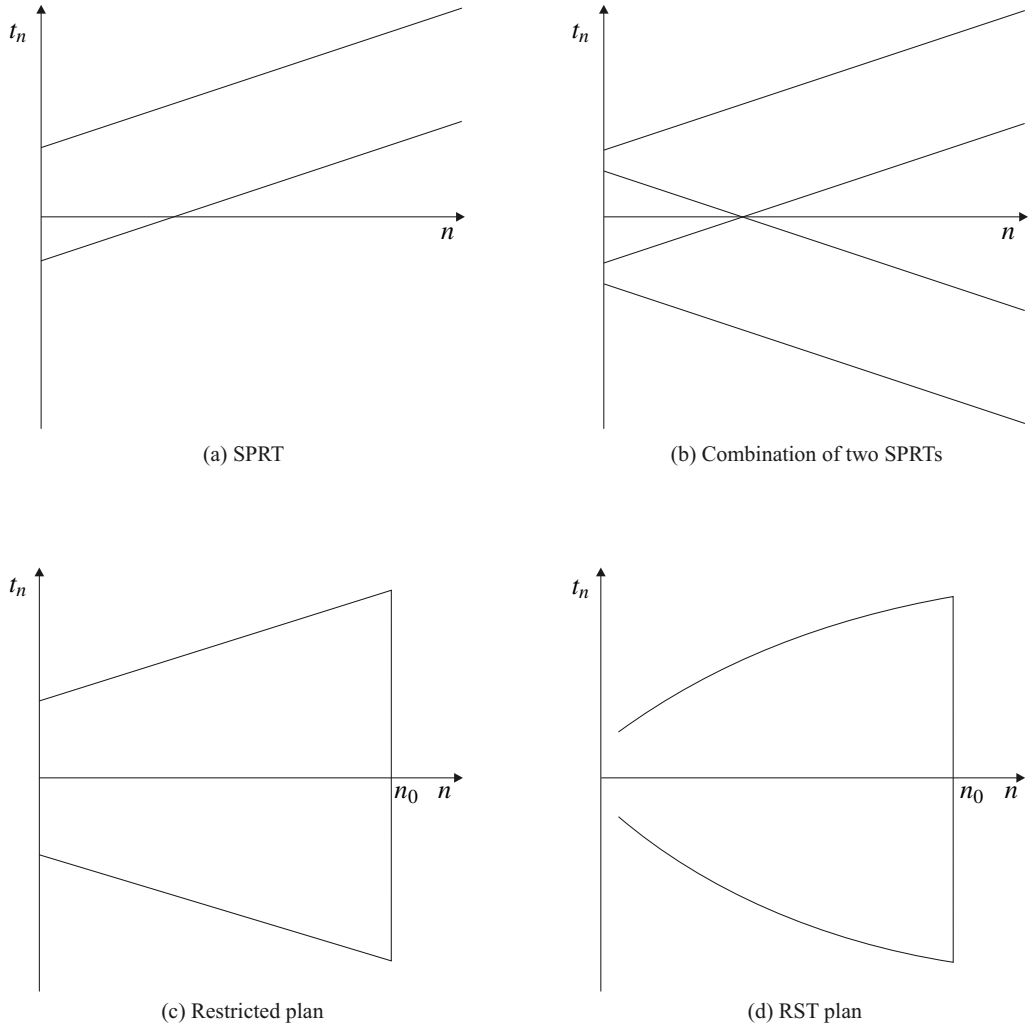
2.1 Early sequential methods

According to Armitage (1990), "[a] scientific investigation is sequential if its conduct at any stage depends on the outcome at previous stages." Probably the earliest application of sequential methods can be found in Dodge and Romig (1929) in which "double sampling schemes" are used in industrial batch sampling for quality monitoring. These two-stage sequential methods were adapted in cancer drug screening trials, e.g. in Gehan (1961), Lee et al. (1979), and Simon (1989). In a theoretical development, Stein (1945) derived a sequential procedure that uses estimated variance from the first-stage sample in choosing the size of the second-stage sample to achieve a desired power of a two-stage t-test.

For a fixed sample test of the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \theta_1$, let $f(x;\theta)$ be the probability density or mass function for a random variable X. According to Neyman and Pearson (1933), one rejects H_0 in favor of H_1 if $L_n > c_\alpha$ where

$$L_n = \prod_{i=1}^{n} \frac{f(x_i;\theta_1)}{f(x_i;\theta_0)}$$

is the likelihood ratio. The critical value c_α is determined for the test to be of size α. Then the test is most powerful, that is, the type II error probability β is smallest amongst all tests with size $\leq \alpha$.



(a) SPRT (b) Combination of two SPRTs

(c) Restricted plan (d) RST plan

Figure 1: *Sequential boundaries from Fig. 6.1 in Armitage (1990).*

Following what came to be known as Neyman-Pearson's fundamental lemma above, Wald (1947) developed the sequential probability ratio test (SPRT) to discriminate between two simple hypotheses. Specifically Wald SPRT shows that when the sample size is not fixed in advance, further improvement is possible. The best procedure in a certain sense made precise by Wald and Wolfowitz (1948) is 1) to continue sampling as long as $B < L_n < A$ for some constant $B < 1 < A$ and 2) to stop sampling and decide in favor of H_1 or H_0 as soon as $L_n > A$ or $L_n < B$, respectively, where

$$A \approx \frac{1-\beta}{\alpha} \text{ and } B \approx \frac{\beta}{1-\alpha}.$$

A specific case when $\theta_0 = 0$ and $\theta_1 > 0$ is a one-tailed test as shown in Fig. 1(a). There are two different versions of its generalization for a two-tailed test with $H_1 : \theta \neq 0$. One is a two-tailed test obtained by defining a density function $f_1 = (f_- + f_+)/2$ where f_- and f_+ are the probability density or mass functions corresponding to alternative hypotheses $H_- : \theta < 0$ and $H_+ : \theta > 0$ in two directions as suggested by Wald (1947) (Chapter 9). The other is a combination of two separate one-tailed tests, each with type I error probability $\alpha/2$, by Sobel and Wald (1949), as shown in Fig. 1(b).

One drawback of SPRTs is that sampling may continue indefinitely. A restricted plan by Armitage (1957) is a modification of the two-tailed version of a SPRT by Sobel and Wald (1949) to avoid this possibility by imposing a maximum sample size with the inner wedge removed or pushed out as shown in Fig. 1(c). A similar sequential plan was later developed by Armitage et al. (1969) as a repeated significance test plan as shown in Fig. 1(d) and described in detail in Subsection 2.3 as a means to adjust the critical value to account for multiple testing leading to a constant critical value. Of note, the operating characteristics of these two sequential tests in Figs. 1(c) and 1(d) are very similar.

2.2 "Sampling to reach a foregone conclusion"

Let $X_1, X_2, \ldots$ be independent and identically distributed and drawn from $N(\mu, \sigma^2)$ with known variance σ^2, and consider a statistical test of $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. For a single sampling plan with a fixed sample size n, one would reject H_0 if and only if $|S_n| > 1.96\sigma\sqrt{n}$ at a significance level $\alpha = 0.05$ where $S_n = \sum_{i=1}^{n} X_i$.

A need for adjustment in the critical value for repeated testing is recognized by the law of the iterated logarithm described here. Assume only that X_i, $i = 1, 2, \ldots$ are simply independent and identically distributed with mean μ and finite variance $0 < \sigma^2 < \infty$. In addition assume that n is not fixed in advance, and data become available sequentially one at a time. If S_n is computed for each $n \geq 1$, $|S_n|$ is certain to exceed $1.96\sigma\sqrt{n}$ for some n, even if H_0 is true, for the law of the iterated logarithm asserts that

$$\limsup_{n\to\infty} \frac{S_n - n\mu}{\sigma\sqrt{2n\log\log n}} = 1 \text{ with probability } 1.$$

Thus an unscrupulous experimenter might be tempted to take a sample of size

$$N = \inf\{n \geq 1 : |S_n| > 1.96\sigma\sqrt{n}\},$$

and report as if it were a fixed sample size and claim rejection of H_0 at a significance level 0.05. However, the experimenter may have to spend some time in the process as the expected sample size under this sampling scheme is $\mathrm{E}(N) = \infty$.

That one can reach a nominal significance by testing repeatedly was aptly described as "sampling to reach a foregone conclusion" by Anscombe (1954).

2.3 Repeated significance tests

Controversy regarding control of type I error probability depending on the approach, be it Bayesian, likelihood-based, or frequentist, led Armitage et al. (1969) to evaluate the type I error probability of the sequential testing procedure described above to settle the score, so to speak. The numerical procedure for computing the type I error probability is described below.

Assume as above that $X_1, X_2, \ldots$ are independent and identically distributed normal random variables with mean μ and, without loss of generality, variance 1. To test H_0 : $\mu = 0$ against $H_1 : \mu \neq 0$ at a significance level α, sampling is terminated the first time when

$$|S_k| > b_k$$

where $b_1, b_2, \ldots$ are boundary values. With the maximum number of observations K, the boundary values have to satisfy the following:

$$\Pr(|S_k| > b_k \text{ for some } k = 1, \ldots, K) = \alpha$$

or equivalently

$$\Pr(|S_1| \leq b_1, \ldots, |S_K| \leq b_K) = 1 - \alpha.$$

The computation of these probabilities can be simplified by noting that f_k, the probability density function of S_k under H_0 in the sequential procedure, satisfies the following recursive definition base on convolution:

$$f_k(s) = \int_{-b_{k-1}}^{b_{k-1}} f_{k-1}(u)\phi(s-u)du \tag{2.1}$$

where f_1 is the standard normal density function ϕ above. This is so because of the independence between S_{k-1} and $S_k - S_{k-1}$, i.e. independent increments in S_k.

With $k^\star$ denoting the random variable for when $|S_k| > b_k$ for the first time, the probability of stopping at or before k is

$$P_k = \Pr(k^\star \le k) = 1 - \Pr(|S_1| \le b_1, \ldots, |S_k| \le b_k) = 1 - \int_{-b_k}^{b_k} f_k(u)du$$

and the probability of stopping at $k^\star = k$, i.e. the exit probability $\Pr(k^\star = k)$, is simply

$$\begin{aligned} P_k - P_{k-1} &= \Pr(|S_1| \le b_1, \ldots, |S_{k-1}| \le b_{k-1}, |S_k| > b_k) \\ &= \int_{-b_{k-1}}^{b_{k-1}} f_{k-1}(u)\{1 - \Phi(b_k - u) + \Phi(-b_k - u)\}du \end{aligned} \tag{2.2}$$

where Φ is the standard normal distribution function. The overall significance of the sequential procedure is determined by

$$\alpha = 1 - \int_{-b_K}^{b_K} f_K(u)du.$$

The recursive definition of f_k above allows direct computation of these probabilities using standard numerical integration methods, e.g. a Newton-Cotes formula of the second order, i.e. Simpson's rule. This same computational procedure works when $\mu \neq 0$ with X_k replaced by $X_k - \mu$. The above observation led to the notion of repeated significance tests as described in Armitage et al. (1969), which in turn paved the way for development of group sequential methods for clinical trials.

2.4 Group sequential methods for clinical trials

Following the seminal work on sequential analysis by Wald (1947), Bross (1952) and Armitage (1954) appear to have been the first to advocate the use of sequential methods in clinical trials. Different from other settings where savings in sample size was the primary motivation for using sequential methods, it was ethical imperatives in clinical trials in considering early termination to avoid unnecessary experimentation on human subjects in the presence of clear evidence of benefits or harms of interventions.

Suppose that response to treatment is a normal random variable with means μ_A and μ_B for treatments A and B, respectively, and known variance σ^2, a typical two-sample problem. Consider a test of $H_0 : \mu_A = \mu_B$ against $H_1 : \mu_A \neq \mu_B$ or, equivalently, $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$ where $\delta = \mu_A - \mu_B$. A fixed sample size test with a significance level $\alpha = 0.05$ with n participants on each treatment rejects H_0 when

$$Z = \left| \frac{\bar{X}_A - \bar{X}_B}{\sqrt{2\sigma^2/n}} \right| > 1.96$$

where $\bar{X}_A$ and $\bar{X}_B$ denote the sample means.

Group sequential designs call for monitoring of accumulating data over time periodically after groups of observations become available using sequential tests. Wald (1947) (pp 101–103) refers to taking groups of observations and applying SPRTs for binary

outcome. One strategy of a group sequential test is to reject the null hypothesis of no treatment difference if, at any of the interim analyses, the test statistic becomes sufficiently large; otherwise, do not reject (accept) the null hypothesis.

Consider examining the accumulating data after a group of every $2n$ observations, n on each treatment, become available, namely,

$$Y_j = \frac{\bar{X}_{Aj} - \bar{X}_{Bj}}{\sqrt{2\sigma^2/n}} \sim N(\delta^\star, 1)$$

where $\delta^\star = \delta/\sqrt{2\sigma^2/n}$, for up to a maximum of K analyses for a maximum of $2nK$ observations. With the score statistics

$$S_k = \sum_{j=1}^{k} Y_j = \sum_{j=1}^{k} \frac{\bar{X}_{Aj} - \bar{X}_{Bj}}{\sqrt{2\sigma^2/n}} \sim N(\delta^\star k, k) \tag{2.3}$$

or the Wald statistics

$$Z_k = S_k/k^{1/2} \sim N(\delta^* k^{1/2}, 1), \tag{2.4}$$

a group sequential test rejects H_0 for the first time when

$$|S_k| > b_k \text{ or equivalently } |Z_k| > c_k.$$

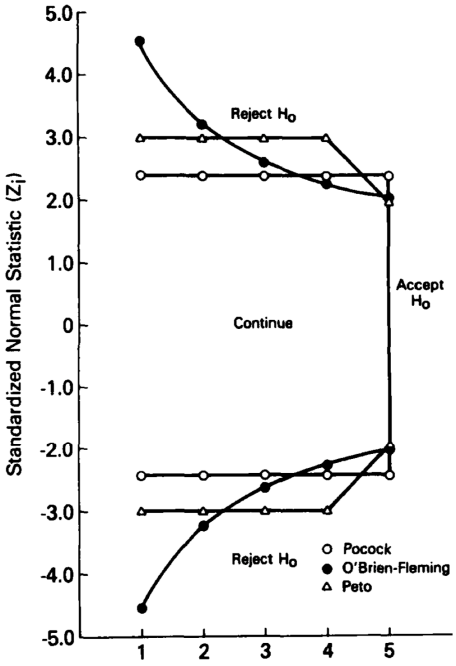


Figure 2: *Group sequential critical values from Fig. 1 in DeMets and Lan (1984).*

Hence, if we want a level α test, we choose the boundary values, $b_1, \ldots, b_K$, or equivalently the critical values, $c_1, \ldots, c_K$, such that, under H_0,

$$\Pr(|S_1| \leq b_1, \ldots, |S_K| \leq b_K) = \Pr(|Z_1| \leq c_1, \ldots, |Z_K| \leq c_K) = 1 - \alpha. \qquad (2.5)$$

Note that there is an equal increment of statistical information in sample size, i.e. $2n$, between sequentially computed test statistics and that the increments are independent so that the computational procedure by Armitage et al. (1969) can be used in this type of group sequential tests.

Several group sequential methods are used for determining the boundary or the critical values. These values for Pocock (1977) and O'Brien and Fleming (1979) group sequential methods are obtained by solving (2.5) under the conditions of $c_1 = \cdots = c_K$ and $b_1 = \cdots = b_K$, respectively (see Fig. 2). Note that Pocock's method is the group sequential version of the repeated significance test method discussed in Subsection 2.3. One practical drawback of these methods is that they depend on the assumption of equal sample size or more generally, equal amount of statistics information, accumulated between two successive analyses. Otherwise the group sequential methods by Pocock (1977) and O'Brien and Fleming (1979) cannot be applied. In order to address this situation, a flexible approach was proposed by Slud and Wei (1982) in which the boundary values, b_k, $k = 1, \ldots K$, are determined with prespecified α_k, $k = 1, \ldots, K$, such that $\alpha_k = P_k - P_{k-1}$ in (2.2) under the null hypothesis and $\sum_{k=1}^{K} \alpha_k = \alpha$, the overall significance level. A practical downside to this approach is the arbitrariness in specifying α_ks and the possibility of the group sequential test not meeting the criterion for early stopping at an interim analysis and meeting the criterion at the next interim analysis with the increment in the statistical information between the two interim analyses in the opposite direction, an obvious logical inconsistency.

Generalizing the idea in Slud and Wei (1982), Lan and DeMets (1983) introduced the notion of "alpha spending" instead of arbitrarily specifying α_ks. As a method of allocating the type I error probability α into α_ks as in Slud and Wei (1982), Lan and DeMets (1983) instead proposed allocating the type I error probability α according to an "error spending function," $\alpha^\star(t)$, which is a nondecreasing function of the information time or fraction t, $0 \leq t \leq 1$, defined below with $\alpha^\star(0) = 0$ and $\alpha^\star(1) = \alpha$. For $k = 1, \ldots, K$, the type I error probability allocated for the k^{th} interim analysis is determined as $\alpha_k = \alpha^\star(t_k) - \alpha^\star(t_{k-1})$ where $t_0 = 0$ and $t_K = 1$ so that $\sum_{k=1}^{K} \alpha_k = \alpha$. For a one-tailed Pocock (1977) and O'Brien and Fleming (1979) procedures, Lan and DeMets (1983) proposed $\alpha^\star_{\text{P}}(t) = \alpha \log\{1 + (e-1)t\}$ and $\alpha^\star_{\text{OF}}(t) = 2\{1 - \Phi(z_{\alpha/2}/\sqrt{t})\}$, respectively, where z_γ is the upper γ quantile of the standard normal distribution. The information fraction t is the fraction of statistical information corresponding to an interim analysis relative to the maximum information required. For example, $t_k = k/K$ for the group sequential tests with equal samples of size n between two successive analyses as in the score statistics in (2.3). If we consider unequal sample sizes n_k between the $(k-1)^{\text{th}}$ and the k^{th} interim analyses, $t_k = n_k/n_K$ instead.

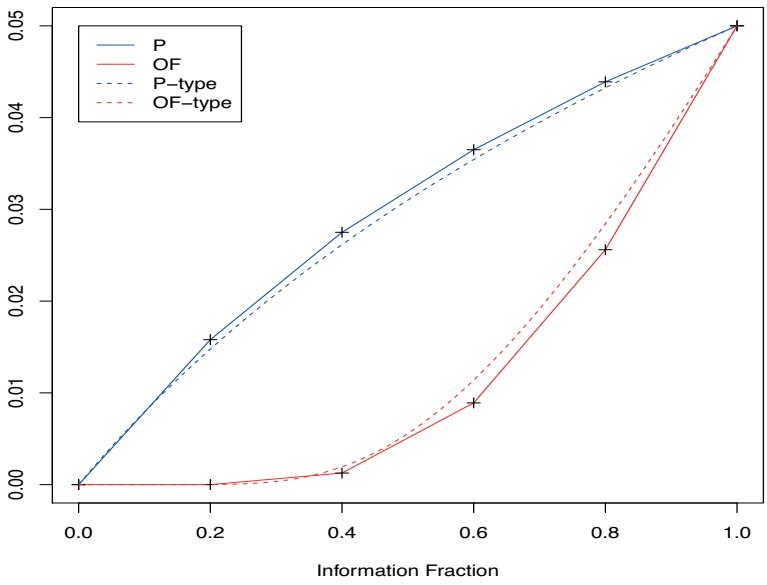


Figure 3: *Cumulative type I error probability for group sequential tests with* $\alpha = 0.05$.

The cumulative type I error probabilities for the Pocock (P) and O'Brien-Fleming (OF) group sequential procedures with $K = 5$ and $\alpha = 0.05$ and the error spending functions $\alpha^{\star}_{\mathrm{P}}(t)$ for Pocock (P-type) and $\alpha^{\star}_{\mathrm{OF}}(t)$ for O'Brien-Fleming (OF-type) from above are plotted in Fig. 3 to indicate similarities between the standard group sequential methods and group sequential methods based on the suitably chosen error spending functions.

From a historical perspective, Pocock (1977), following the repeated significance test of Armitage et al. (1969), popularized the group sequential methods for clinical trials with normal outcome. However, it was Elfring and Schultz (1973) who first coined the term "group sequential designs" for clinical trials with binary outcome. Jennison and Turnbull (1990) present a detailed review of group sequential methods including comparisons of methods by Pocock (1977), O'Brien and Fleming (1979), Slud and Wei (1982), and Lan and DeMets (1983).

2.5 Covariance under independent increments

As noted earlier in Subsection 2.3, in order to apply group sequential methods, one has to solve the following multivariate integral

$$\int_{-b_1}^{b_1} \cdots \int_{-b_K}^{b_K} f(s_1, \ldots, s_K) ds_1 \cdots ds_K = 1 - \alpha$$

where f is the joint density function of the sequential test statistics. However, if the following holds

$$\text{Cov}(S_k, S_l) = \text{Var}(S_k) \text{ or equivalently } \text{Cov}(S_{k-1}, S_k - S_{k-1}) = 0$$

for $1 \le k \le l \le K$ with $S_0 = 0$, i.e. if the sequential test statistics have independent increments, the multivariate integration above becomes univariate integration involving simple recursion based on convolution as indicated in (2.1).

To assess the joint distributions of S_k in (2.3) or Z_k in (2.4), $1 \le k \le K$, consider the fully sequential setting again as in Subsection 2.3. From the standard normal theory and the independent increments structure of S_k, it follows that the joint distribution of the score statistics S_k, $1 \le k \le K$, is multivariate normal with marginals $S_k \sim N(\mu k, k)$ and covariance

$$\text{Cov}(S_k, S_l) = k = \text{Var}(S_k),\ 1 \le k \le l \le K.$$

Since S_k is equivalent to the Wald statistic $Z_k = S_k/k^{1/2} = S_k/\sqrt{\text{Var}(S_k)}$, the corresponding joint distribution of the Wald statistics Z_k, $1 \le k \le K$, are found to be multivariate normal with marginals $Z_k \sim N(\mu k^{1/2}, 1)$ and

$$\text{Cov}(Z_k, Z_l) = \sqrt{k/l} = \sqrt{\text{Var}(S_k)/\text{Var}(S_l)},\ 1 \le k \le l \le K.$$

Hence, any one of the two conditions above gives an independent increments structure of the sequential test statistics. For the two-sample group sequential test as described in Subsection 2.4, replacing μ with δ^*, these results also hold.

More generally, three different test statistics can be considered as in Jennison and Turnbull (1997). For $1 \le k \le l \le K$, the following holds:

$$\hat{\theta}_k \overset{a}{\sim} N(\theta, \mathcal{I}_k^{-1}(\theta)) \text{ and } \text{Cov}(\hat{\theta}_k, \hat{\theta}_l) = \text{Var}(\hat{\theta}_l) = \mathcal{I}_l^{-1}(\theta) \tag{2.6}$$

for the maximum likelihood estimates where $\mathcal{I}_k(\theta)$ is the Fisher information;

$$S_k \overset{a}{\sim} N(\theta \mathcal{I}_k(\theta), \mathcal{I}_k(\theta)) \text{ and } \text{Cov}(S_k, S_l) = \text{Var}(S_k) = \mathcal{I}_k(\theta) \tag{2.7}$$

for the score statistics; and

$$Z_k \overset{a}{\sim} N(\theta \mathcal{I}_k^{1/2}(\theta), 1) \text{ and } \text{Cov}(Z_k, Z_l) = \sqrt{\mathcal{I}_k(\theta)/\mathcal{I}_l(\theta)} \tag{2.8}$$

for the Wald statistics $Z_k = \hat{\theta}_k/\text{SE}(\hat{\theta}_k)$ where SE stands for standard error.

Note that these distributional properties of the sequential test statistics are still true under the general alternatives as well as the null hypothesis, and hence power of the sequential tests can also be evaluated through the univariate integration technique as in McPherson and Armitage (1971). When the underlying distribution is not normal, we consider a class of local alternatives $\{\mu_n\}$, where $\sqrt{n}\mu_n \to \delta \ne 0$. Then normality and an independent increments structure of the sequentially computed test statistics can be established asymptotically under the null and a class of local alternatives so that the standard sequential procedures described in this section are still applicable asymptotically.

2.6 Intuition about independent increments

With normal outcome, it is intuitive that group sequential test statistics would have independent increments, thus allowing application of the classical group sequential methods. With time to event outcome, it is unclear since each participant contributes follow-up data possibly multiple times over group sequential tests. With longitudinal outcome, again it is unclear since each participant contributes follow-up data multiple times longitudinally. Both with longitudinal data and failure time data, a participant contributes data more than once over the course of study in group sequential tests and as a consequence it is not intuitive why sequential tests statistics would have independent increments.

As summarized in Jennison and Turnbull (1990), independent increments structures have been found to hold in many circumstances case by case. Scharfstein et al. (1997) showed with great generality that the efficient score statistics in parametric and semiparametric models have an independent increments structure. Jennison and Turnbull (1997) also gave a unified explanation based on efficiency of the test statistics for the independent increments structure. For instance, in our fully sequential setting, since the sample mean $\bar{X}_k$ is the maximum likelihood estimator, or least squares estimator of μ, the corresponding sequential score and Wald tests, S_k and Z_k, have an independent increments structure following their theorems. In this review paper, we consider the group sequential score tests with independent increments derived from several estimating methods such as the maximum likelihood and least squares method. For some of them, the independent increments structures are explained by efficiency of the test statistics, while it is not for others.

3 Joint distributions of sequential test statistics

In this section we provide a review of the historical development of independent increments in group sequential tests used in clinical trials with longitudinal data and failure time data as the primary endpoint of interest for evaluation of efficacy of intervention. The emphasis on these types of outcome data is because of the fact that they are widely used in clinical trials in chronic diseases. But more importantly it is not intuitive as to why some group sequential tests for these types of outcome data have an independent increments structure while others do not. This is in contrast to the settings in which outcome data are measured only once from each participant, which intuitively have an independent increment structure.

3.1 Longitudinal data

The joint distribution of sequential test statistics for longitudinal data has been investigated by many authors for application of group sequential methods in clinical trials with such outcome data: Armitage et al. (1985), Geary (1988), Lee and DeMets (1991), Reboussin et al. (1992), and Wu and Lan (1992) based on parametric models; Lee and

DeMets (1992) based on linear rank tests; Su and Lachin (1992) based on a multivariate generalization of the Hodges and Lehmann (1963) estimator of a location shift; Wei et al. (1990a), Gange and DeMets (1996), and Lee et al. (1996) based on semiparametric models in generalized estimation equations; and Spiessens et al. (2002) based on a random-effects model for longitudinal ordinal outcome. Lee (1994) and Spiessens et al. (2000) provide review of some of these sequential tests for longitudinal data.

When the primary outcome is longitudinal data with repeated measurements, each participant can contribute outcome data to test statistics more than once. Thus it is not intuitively obvious that sequential test statistics can have independent increments due to apparent correlation among repeated measurements from the same participant. Indeed the joint distributions of the sequential test statistics by Armitage et al. (1985), Geary (1988), Wei et al. (1990a), Lee and DeMets (1992), and Su and Lachin (1992), all turn out to have correlated increments. But as summarized below, properly formulated test statistics and semiparametric-efficient tests for longitudinal data under various parametric and semiparametric models have independent increments.

Under a linear mixed-effects model of Laird and Ware (1982), Lee and DeMets (1991) show that the asymptotic joint distribution of the sequential test statistics for comparing the rates of change computed over time is multivariate normal under missing at random and includes as special cases those by Armitage et al. (1985) and Geary (1988). Later Reboussin et al. (1992) showed that the test statistics of Lee and DeMets (1991) have an independent increments structure.

In order to account for informative drop-out, Wu and Lan (1992) proposed group sequential tests to compare areas under the response change curves between two treatments based on the two-stage random effects model of Wu and Bailey (1989). It is shown that when the response curve is linear and drop-out non-informative, the test by Wu and Lan (1992) reduces to that by Lee and DeMets (1991) above and that the joint distribution of the test statistics computed over time has independent increments.

Wei et al. (1990a), Gange and DeMets (1996), and Lee et al. (1996) all proposed a group sequential test based on a semiparametric model using the generalized estimating equations approach of Liang and Zeger (1986). Wei et al. (1990a) assume an independence model for the working variance for repeated measures, while Gange and DeMets (1996) and Lee et al. (1996) assume that the covariance matrix for repeated measures is correctly specified or consistently estimated by the working covariance matrix as in Liang et al. (1992).

As indicated by Scharfstein et al. (1997), the joint distribution of the sequentially computed score statistics based on an independence model by Wei et al. (1990a) results in correlated increments as the test is not semiparametric efficient. Gange and DeMets (1996) show that the joint distribution of the regression estimators, i.e. estimators based on the generalized estimating equations, over time is asymptotically multivariate normal with independent increments, while Lee et al. (1996) show that the joint distributions of the sequentially computed score and Wald statistics both are asymptotically multivariate normal with independent increments.

As noted above, standard group sequential methods can be used if one uses an efficient test statistics over time. With random-effects models for ordinal longitudinal data, a Wald-type test can be used with standard group sequential methods. Spiessens et al. (2002) show that, even when the random-effects distribution is misspecified, the joint distribution of the Wald-type test computed over time is asymptotically multivariate normal and showed through simulation studies that a sandwich-type correction to the covariate matrix leads to an approximately independent increments structure.

3.2 Failure time data

Many authors also investigated the joint distribution of sequential test statistics for failure time data under various settings for application of group sequential methods: Tsiatis (1981) and Sellke and Siegmund (1983) under the proportional hazards model; Gail et al. (1982) for two-sample logrank score test; Tsiatis (1982), Slud (1994), and Gu and Lai (1991) for general linear rank tests; Slud and Wei (1982) for the modified Wilcoxon statistics, i.e. a generalized Wilcoxon test by Gehan (1965); Tsiatis et al. (1985) and Gu and Ying (1995) under the proportional hazards model with covariate adjustment; Lin (1992) for logrank tests adjusting for covariates under the accelerated failure time model; and Tsiatis et al. (1995) for general parametric survival models.

When failure time is a primary outcome, each participant can contribute statistical information to group sequential tests more than once before event of interest or random censoring occurs. Hence it seems natural for the increments in successive test statistics to be correlated. Indeed the joint distributions of the test statistics over time by Slud and Wei (1982) for Gehan's test by Gehan (1965) and by Lin (1992) for the logrank test under the accelerated failure time model turn out to have correlated increments. In the case of a general class of linear rank tests, Tsiatis (1982) provides the condition for the weight function under which the joint distribution of the linear rank tests computed over time has independent increments.

Tsiatis (1981) was the first to develop the joint distribution of sequential test statistics and establish independent increments for a sequential test for failure time data. First the asymptotic joint distribution of the sequentially computed score statistics for the proportional hazards model was established and shown to to converge asymptotically to a multivariate Gaussian process with independent increments when participants enter randomly throughout the course of the trial. This allows group sequential methods to be based on the logrank test as a special case of the efficient scores test for the proportional hazards model in clinical trials with failure time data subject to random censoring, thus proving the conjecture made earlier in Armitage (1975) (pp 140–143).

Gail et al. (1982) investigated the operating characteristics of the logrank score test, computed after fixed numbers of events and applied to various group sequential methods, using simulation studies. They show empirically that the joint distribution of the logrank score test computed over time follows a multivariate normal distribution with independent increments reasonably well in a realistic setting in clinical trials.

Tsiatis (1982) generalizes the results in Tsiatis (1981) to a general class of nonparametric linear rank tests statistics and shows that the asymptotic joint distribution of the sequential test statistics within this general class of nonparametric tests is a multivariate normal distribution. This general class of nonparametric tests is characterized by a random function corresponding to the weight functions described by Tarone and Ware (1977) and Prentice and Marek (1979) and as a special case includes a constant weight for the logrank test, a weight function for the modified Wilcoxon test which is the survival function.

Sellke and Siegmund (1983) show that the score process of the partial likelihood and the maximum partial likelihood estimator under the proportional hazards model behave asymptotically like a Brownian motion. This relies on the approximation of the score process by a suitable martingale and a random rescaling of time based on the observed Fisher information. As such, the resulting joint distributions of the score process and the maximum partial likelihood estimator over time both have independent increments.

Slud (1994) shows that under the null hypothesis of no difference in survival distributions the sequentially computed logrank statistics of Mantel (1966) have exactly uncorrelated increments under very general patterns of enrollment, allocation to treatment and lost to follow-up in clinical trials. Gu and Lai (1991) considers the general class of linear rank test statistics investigated in Tsiatis (1982) and develops a general weak convergence theory for the joint distribution of the sequential linear rank test statistics for two sample problems in a realistic clinical trial setting.

Tsiatis et al. (1985) investigates the joint distribution of the sequentially computed efficient scores for the treatment effect derived from a partial likelihood under the proportional hazards model with adjustment for other covariates. They show that the sequential efficient scores test for the treatment effect in the presence of other covariates has asymptotically the same joint distribution as the sequentially computed ordinary logrank test with no covariates. The motivation for this work was the efficiency gain in the test by adjusting for the effects of other covariates. Gu and Ying (1995) show that a general Cox-type partial likelihood score process for staggered entry with covariate adjustment is asymptotically equivalent to a Gaussian process with independent increments, including the case in which the covariates being adjusted for are not independent of the covariates of primary interest, typically a randomized treatment indicator.

Tsiatis et al. (1995) consider the joint distribution of sequentially computed score statistics and the maximum likelihood estimator in parametric models for failure time data in the presence of nuisance parameters. By representing the sequentially computed score test as a stochastic integral of a counting process martingale, they drive the asymptotic joint distribution of the test statistics over time and show that the joint distributions of the score test and the maximum likelihood estimator are multivariate normal with independent increments. This work and the work by Lee et al. (1996) served as a seed for group sequential methods based on semiparametric efficient test statistics by Scharfstein et al. (1997).

Scharfstein et al. (1997) noted that joint distributions of many group sequential statistics used to analyze longitudinal or failure time data are asymptotically multivariate normal with an independent increments structure. This limiting distribution arises naturally when one uses an efficient test statistic to test a single parameter in a semiparametric model. They develop most general results based on semiparametric efficient tests and show that many previously developed cases of independent increments structure are a special case of a semiparametric efficient test.

4 Independent increments

In this section we review most general cases of independent increments for sequential tests for longitudinal and failure time data. First we define some notations and consider the formulation of the problem.

Consider a group sequential study with a maximum number of K interim analyses at calendar times t_k, $k = 1,\ldots,K$. We allow staggered entry of subjects and denote n_k to be the number of subjects who have entered the study at the k^{th} interim analysis. Let Y_{ik} be the outcome of the ith subject. When repeated measures are made as in a longitudinal study, let $Y_{ik} = (Y_{i1k},\ldots,Y_{i,d_{ik},k})^{\top}$ where d_{ik} denote the number of repeated measures of the ith subjects. At each k, Y_{ik}, $i = 1,\ldots,n_k$, are assumed to be independent. Let $X_{ik} = (Z_{ik}, W_{ik})$ denote a $d_{ik} \times p$ dimensional covariate (design) matrix including a treatment indicator Z_{ik} and $p-1$ time-varying covariate vectors W_{ik}, and let $\theta = (\gamma, \beta^{\top})^{\top}$ denote a corresponding parameter vector which consists of a treatment effect parameter γ and covariate effect parameters β. The total number of subjects at the last analysis is set as $n_K = n$, and let T_i be the entry time of the ith subject.

Our primary interest is focused on the group sequential tests with independent increments for the hypotheses of

$$H_0 : \gamma = 0 \text{ vs } H_1 : \gamma \neq 0 \tag{4.1}$$

where the parameters β are regarded as nuisance parameters adjusting for covariates.

A test for the hypotheses in (4.1) is obtained from the "score" vector. At the k^{th} interim analysis, let the p dimensional score vector or, more generally, "estimating equations" to be used to estimate θ, be denoted by

$$S_k(\theta) = \sum_{i=1}^{n_k} S_{ik}(\theta), \tag{4.2}$$

and let $\hat{\theta}_k$ denote the estimator of θ satisfying $S_k(\hat{\theta}_k) = 0$ if it exists. For example, in the fully sequential method described in Subsection 2.3, we can consider a kind of score vector $S_k(\mu) = \sum_{i=1}^{k}(X_i - \mu)$. By solving the estimating equation $S_k(\hat{\mu}_k) = 0$, it produces the estimator $\hat{\mu}_k = \bar{X}_k$ and the Wald test $\hat{\mu}_k/\mathrm{SE}(\hat{\mu}_k) = S_k/\sqrt{k}$, where SE stands for standard error. Note that, under the null hypothesis of $\mu = 0$, the score vector becomes the score test $S_k = S_k(0)$ which is equivalent to the Wald test.

In fact, the score vector given by (4.2) contains several important estimating equation vectors such as the efficient score vector in (4.11) defined by differentiating a log-likelihood with respect to θ and the "least squares" score vector (4.5) obtained from the least squares estimation method. In the sequel, the explicit form of score vectors will be defined case by case.

To construct a sequential score statistics in the presence of nuisance parameters β, we partition, under the null hypothesis of $\gamma = 0$, the score vector (4.2) as

$$S_k(\theta)|_{\gamma=0} = (S_{k,\gamma}(\beta), S_{k,\beta}(\beta)^{\mathsf{T}})^{\mathsf{T}}$$

where $S_{k,\gamma}(\beta)$ denotes a score function with respect to the treatment effects parameter γ and $S_{k,\beta}(\beta)$ denotes a $(p-1)$ dimensional score vector with respect to the nuisance parameters β. Then as test statistics at the k^{th} interim analysis, one can use the score statistics $S_{k,\gamma}(\hat{\beta}_k)$ and Wald statistics $\hat{\gamma}_k/\text{SE}(\hat{\gamma}_k)$ where $\hat{\beta}_k$ is the restricted estimator of β computed under the null hypothesis and $\hat{\gamma}_k$ is the estimator of γ obtained by solving $S_k(\hat{\theta}_k) = 0$. Though both the Wald and the score tests can be used to test the null hypothesis, we will use mainly the score tests for convenience.

The score statistics $S_{k,\gamma}(\hat{\beta}_k)$ are usually expressed, at least approximately, as a linear combination of the scores $S_{k,\gamma}(\beta)$ and $S_{k,\beta}(\beta)$ so that the joint distribution and the independent increment structure of the sequentially computed score statistics can be established by the distributional properties of $S_{k,\gamma}(\beta)$ and $S_{k,\beta}(\beta)$. For example, the Taylor expansions of $S_{k,\gamma}(\hat{\beta}_k)$ and $S_{k,\beta}(\hat{\beta}_k)$ at $\beta = \beta_0$, when applicable, yield

$$S_{k,\gamma}(\hat{\beta}_k) \simeq S_{k,\gamma}(\beta_0) + S'_{k,\gamma}(\beta_0)(\hat{\beta}_k - \beta_0),$$

$$0 = S_{k,\beta}(\hat{\beta}_k) \simeq S_{k,\beta}(\beta_0) + S'_{k,\beta}(\beta_0)(\hat{\beta}_k - \beta_0),$$

where $S'_{k,\gamma}(\beta_0) = \partial S_{k,\gamma}(\beta)/\partial\beta|_{\beta=\beta_0}$ and $S'_{k,\beta}(\beta_0) = \partial S_{k,\beta}(\beta)/\partial\beta|_{\beta=\beta_0}$. They are combined yielding

$$S_{k,\gamma}(\hat{\beta}_k) \simeq S_{k,\gamma}(\beta_0) - S'_{k,\gamma}(\beta_0)\{S'_{k,\beta}(\beta_0)\}^{-1} S_{k,\beta}(\beta_0).$$

Since the score vector (4.2) depends only on observations accumulated up to stage k and it has a form of sum of independent observations, so do $S_{k,\gamma}(\beta)$ and $S_{k,\beta}(\beta)$. Even in the case of repeated measurement ($d_{ik} \geq 2$), we can define $S_{ik}(\theta)$ to accomodate the dependency in Y_{ik} through such a method used in the generalized least squares estimation, and hence the structure of sum of independent observations will still hold. Therefore, applying the central limit theorem, the joint distribution of the sequential score statistics $S_{k,\gamma}(\hat{\beta}_k)$ as well as those of $S_{k,\gamma}(\beta)$ and $S_{k,\beta}(\beta)$ would be a (asymptotic) multivariate normal distribution under some regularity conditions, and also they are expected to have the independent increment structure. If it is the case, the standard group sequential methods described in Subsection 2.4 can be applied to the score statistics $S_{k,\gamma}(\hat{\beta}_k)$ to carry out testing for the null hypothesis.

In the asymptotic approach, to avoid the problem caused when n_k are random, we assume the data structure described in Scharfstein et al. (1997). That is, at the k^{th} interim analysis, consider the accumulated data set $\{Y_{ik}, i = 1,\ldots,n_k\}$ as $\{(Y_{ik}, I(T_i \leq t_k)), i = 1,\ldots,n\}$ where $I(T_i \leq t_k)$ is defined as 1 if the ith patient has entered the study by the time of the k^{th} interim analysis and 0 otherwise. Then the score vector (4.2) can be written as

$$S_k(\theta) = \sum_{i=1}^{n} S_{ik}(\theta) I(T_i \leq t_k), \tag{4.3}$$

and we can establish the asymptotic results based on the total sample size n. With this in mind, we will use the expression of (4.2) rather than that of (4.3).

The more detailed theory of maximum likelihood and generalized least squares estimation can be found, for example, in Cox and Hinkley (1974) and McCullagh and Nelder (1989), respectively.

4.1 Parametric regression models for independent data

We start with the simple model for independent data. Consider a regression model below:

$$\begin{aligned} Y_{ik} &= X_{ik}\theta + \epsilon_{ik} \\ &= Z_{ik}\gamma + W_{ik}\beta + \epsilon_{ik},\ i = 1,\ldots,n_k\ ;\ k = 1,\ldots,K, \end{aligned} \tag{4.4}$$

where the independent error terms ϵ_{ik} have a common distribution function F and a common density function f with mean zero and variance σ^2. Then the usual least squares score vector, at the k^{th} interim analysis, is defined by

$$S_k(\theta) = \sum_{i=1}^{n_k} X_{ik}^{\mathsf{T}}(Y_{ik} - X_{ik}\theta). \tag{4.5}$$

From (4.5), the partitioned scores of $S_k(\theta)$ under the null hypothesis of $\gamma = 0$ in the presence of a nuisance parameter β are given by

$$S_{k,\gamma}(\beta) = \sum_{i=1}^{n_k} Z_{ik}(Y_{ik} - W_{ik}\beta) \tag{4.6}$$

and

$$S_{k,\beta}(\beta) = \sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}}(Y_{ik} - W_{ik}\beta). \tag{4.7}$$

Under the null hypothesis, the restricted estimator of β satisfying $S_{k,\beta}(\hat{\beta}_k) = 0$ in (4.7) is the least squares estimator denoted by

$$\hat{\beta}_k = \left(\sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} W_{ik} \right)^{-1} \sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} Y_{ik}.$$

Plugging it into (4.6), the score statistics $S_{k,\gamma}(\hat{\beta}_k)$ are written as a linear combination of observations Y_{ik} as follows:

$$\begin{aligned} S_{k,\gamma}(\hat{\beta}_k) &= \sum_{i=1}^{n_k} Z_{ik}(Y_{ik} - W_{ik}\hat{\beta}_k) \\ &= \sum_{i=1}^{n_k} Z_{ik} Y_{ik} - \left(\sum_{i=1}^{n_k} Z_{ik} W_{ik} \right) \left(\sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} W_{ik} \right)^{-1} \sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} Y_{ik}. \end{aligned} \tag{4.8}$$

Note that we can also express the score statistics (4.8) as one having a form of (4.6),

$$S_{k,\gamma}(\hat{\beta}_k) = S_{k,\gamma}(0) - S'_{k,\gamma}(0)\{S'_{k,\beta}(0)\}^{-1} S_{k,\beta}(0)$$

or equivalently,

$$S_{k,\gamma}(\hat{\beta}_k) = S_{k,\gamma}(0) - \Gamma_{k,\gamma\beta}\Gamma_{k,\beta\beta}^{-1} S_{k,\beta}(0) \tag{4.9}$$

where $\Gamma_{k,\gamma\beta}$ and $\Gamma_{k,\beta\beta}$ are submatrices of the partitioned matrix

$$\Gamma_k = \mathrm{Var}\{(S_{k,\gamma}(0), S_{k,\beta}(0)^{\mathsf{T}})^{\mathsf{T}}\} = \begin{bmatrix} \Gamma_{k,\gamma\gamma} & \Gamma_{k,\gamma\beta} \\ \Gamma_{k,\beta\gamma} & \Gamma_{k,\beta\beta} \end{bmatrix}. \tag{4.10}$$

From the equation (4.8), (4.9) and (4.10), we have

$$E\{S_{k,\gamma}(\hat{\beta}_k)\} = \gamma I_k$$

and

$$\mathrm{Var}\{S_{k,\gamma}(\hat{\beta}_k)\} = I_k$$

where

$$I_k = \Gamma_{k,\gamma\gamma} - \Gamma_{k,\gamma\beta}\Gamma_{k,\beta\beta}^{-1}\Gamma_{k,\beta\gamma} = \left\{ \sum_{i=1}^{n_k} Z_{ik}^2 - \left(\sum_{i=1}^{n_k} Z_{ik} W_{ik} \right) \left(\sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} W_{ik} \right)^{-1} \left(\sum_{i=1}^{n_k} Z_{ik} W_{ik} \right)^{\mathsf{T}} \right\} \sigma^2.$$

To show the independent increments structure of $S_{k,\gamma}(\hat{\beta}_k)$, we first express $S_{l,\gamma}(0)$ and $S_{l,\beta}(0)$ in the equations (4.9) as sums of two independent variables,

$$S_{l,\gamma}(0) = S_{k,\gamma}(0) + \{S_{l,\gamma}(0) - S_{k,\gamma}(0)\}$$

and

$$S_{l,\beta}(0) = S_{k,\beta}(0) + \{S_{l,\beta}(0) - S_{k,\beta}(0)\}$$

for $k \leq l$. Then we can show that

$$\mathrm{Cov}\{S_{k,\gamma}(0), S_{l,\gamma}(0)\} = \mathrm{Var}\{S_{k,\gamma}(0)\} = \Gamma_{k,\gamma\gamma},$$

$$\mathrm{Cov}\{S_{k,\beta}(0), S_{l,\beta}(0)\} = \mathrm{Var}\{S_{k,\beta}(0)\} = \Gamma_{k,\beta\beta}$$

$$\mathrm{Cov}\{S_{k,\gamma}(0), S_{l,\beta}(0)\} = \mathrm{Cov}\{S_{l,\gamma}(0), S_{k,\beta}(0)\} = \mathrm{Cov}\{S_{k,\gamma}(0), S_{k,\beta}(0)\} = \Gamma_{k,\gamma\beta},$$

and

$$\mathrm{Var}\{S_{l,\beta}(0)\} = \mathrm{Var}\{S_{k,\beta}(0)\} = \Gamma_{k,\beta\beta}.$$

These equations produce the independent increments such that

$$\mathrm{Cov}\{S_{k,\gamma}(\hat{\beta}_k), S_{l,\beta}(\hat{\beta}_l)\} = I_k = \mathrm{Var}\{S_{k,\gamma}(\hat{\beta}_k)\}.$$

We established the independent increments structure of sequentially computed score statistics $S_{k,\gamma}(\hat{\beta}_k)$ without normality assumption for the error distribution. Hence, one might construct the exact sequential tests by replacing the normal density function with an underlying density function f in the methods given in Subsection 2.3. If the asymptotic methods are preferred for a non-normal distribution, we can use the asymptotic results established by the multivariate central limit theorem and the Cramér-Wold device. That is, the asymptotic joint distribution of the sequential score statistics $n^{-1/2}S_{k,\gamma}(\hat{\beta}_k)$, $k = 1, \ldots, K\}$, under the null hypothesis, is multivariate normal with mean 0 and covariance matrix

$$\mathrm{Cov}_A\{n^{-1/2}S_{k,\gamma}(\hat{\beta}_k), n^{-1/2}S_{l,\beta}(\hat{\beta}_l)\} = \mathrm{Var}_A\{n^{-1/2}S_{k,\gamma}(\hat{\beta}_k)\} = \bar{I}_k, \;\; 1 \leq k \leq l \leq K,$$

where Cov_A and Var_A denote asymptotic covariance and variance matrices and

$$\bar{I}_k = \lim_{n\to\infty} n^{-1} I_k.$$

Further, under a class of local alternatives $\{\gamma_n\}$, where $\sqrt{n}\gamma_n \to \delta \neq 0$, we can show that the asymptotic distribution of $n^{-1/2}S_{k,\gamma}(\hat{\beta}_k)$ is normal with mean $\delta\bar{I}_k$ and the same variance as under the null hypothesis.

It should be mentioned that the variance σ^2 has been assumed known. In addition to the known variance case, the asymptotic results are still valid if there exists a consistent estimator of the variance when unknown. Although there are some exact tests such as exact t, χ^2 and F tests proposed by Jennison and Turnbull (1991), we restrict our attention to two cases: the known variance case and the unknown variance case where a consistent estimator exists.

For the regression model (4.4), the score vector (4.5) coincides with the efficient score vector based on the likelihood function when the underlying distribution is normal.

As shown in Jennison and Turnbull (1997) and Scharfstein et al. (1997), the asymptotic joint distribution of the sequentially computed efficient score statistics is multivariate normal with independent increments for more general models.

To summarize their results, we consider the model given in Jennison and Turnbull (1997) where Y_{ik} has a density function $f_{ik}(y_{ik};\theta)$ satisfying some regularity conditions necessary to establish the asymptotic results. Then, for observation i, defining the efficient score $S_{ik}(\theta)$ and information matrix I_{ik} as

$$S_{ik}(\theta) = \frac{\partial}{\partial\theta}\log f_{ik}(Y_{ik};\theta) \tag{4.11}$$

and

$$I_{ik}(\theta) = E\left\{-\frac{\partial}{\partial\theta}S_{ik}(\theta)^{\mathsf{T}}\right\},$$

we have, at the k^{th} interim analysis, the efficient score vector $S_k(\theta) = \sum_{i=1}^{n_k} S_{ik}(\theta)$ and information matrix $I_k(\theta) = \sum_{i=1}^{n_k} I_{ik}(\theta)$. Note that $I_k(\theta) = \text{Var}\{S_k(\theta)\}$. Further, taking $\hat{\beta}_k$ as the restricted maximum likelihood estimator of β under the null hypothesis, the efficient score statistics $S_{k,\gamma}(\hat{\beta}_k)$ can be approximated as, for a fixed β_0 of β,

$$S_{k,\gamma}(\hat{\beta}_k) \simeq S_{k,\gamma}(\beta_0) - I_{k,\gamma\beta}I_{k,\beta\beta}^{-1}S_{k,\beta}(\beta_0)$$

where $I_{k,\gamma\beta}$ and $I_{k,\beta\beta}$ are submatrices of the partitioned matrix

$$I_k\{(0,\beta_0')'\} = \text{Var}\{(S_{k,\gamma}(\beta_0), S_{k,\beta}(\beta_0)^{\mathsf{T}})^{\mathsf{T}}\} = \begin{bmatrix} I_{k,\gamma\gamma} & I_{k,\gamma\beta} \\ I_{k,\beta\gamma} & I_{k,\beta\beta} \end{bmatrix}.$$

Therefore, by applying the same arguments as those for the least squares method, it can be shown that the asymptotic joint distribution of the sequential score statistics $n^{-1/2}S_{k,\gamma}(\hat{\beta}_k), k = 1,\ldots,K\}$, is multivariate normal with mean μ and covariance matrix

$$\text{Cov}_A\{n^{-1/2}S_{k,\gamma}(\hat{\beta}_k), n^{-1/2}S_{l,\beta}(\hat{\beta}_l)\} = \bar{I}_k, \;\; 1 \le k \le l \le K,$$

where $\bar{I}_k = \lim_{n\to\infty} n^{-1}(I_{k,\gamma\gamma} - I_{k,\gamma\beta}I_{k,\beta\beta}^{-1}I_{k,\beta\gamma})$ and μ is 0 under the null hypothesis and $\delta\bar{I}_k$ under local alternatives. The variance matrix $\bar{I}_k$ can be replaced by the consistent estimator based on sample information matrices

$$\hat{I}_{ik}\{(0,\hat{\beta}_k^{\mathsf{T}})^{\mathsf{T}}\} = -\frac{\partial}{\partial\theta}S_{ik}(\theta)^{\mathsf{T}}\Big|_{\gamma=0,\beta=\hat{\beta}_k}.$$

4.2 Longitudinal data

In this subsection, we review selected recently developed methods for group sequential tests which, when properly formulated, turn out to have independent increments, starting with parametric models followed by semiparametric models. We still consider the regression model (4.4) and discuss methods based on the generalized least squares estimates and generalized estimating equations rather than the maximum likelihood estimates.

4.2.1 Parametric regression models

For the model (4.4), assume $d_{ik} \geq 1$ and ϵ_{ik} has mean 0 and variance matrix V_{ik}. Then, based on the generalized least squares methods, the score vector $S_k(\theta)$, the generalized least squares estimator $\hat{\beta}_k$ of β under the null hypothesis and score statistics $S_{k,\gamma}(\hat{\beta}_k)$ are given by

$$S_k(\theta) = \sum_{i=1}^{n_k} X_{ik}^{\mathsf{T}} V_{ik}^{-1} (Y_{ik} - X_{ik}\theta), \tag{4.12}$$

$$\hat{\beta}_k = \left(\sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} V_{ik}^{-1} W_{ik} \right)^{-1} \sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} V_{ik}^{-1} Y_{ik},$$

and

$$\begin{aligned} S_{k,\gamma}(\hat{\beta}_k) &= \sum_{i=1}^{n_k} Z_{ik}^{\mathsf{T}} V_{ik}^{-1} (Y_{ik} - W_{ik}\hat{\beta}_k) \\ &= \sum_{i=1}^{n_k} Z_{ik}^{\mathsf{T}} V_{ik}^{-1} Y_{ik} - \left(\sum_{i=1}^{n_k} Z_{ik}^{\mathsf{T}} V_{ik}^{-1} W_{ik} \right) \left(\sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} V_{ik}^{-1} W_{ik} \right)^{-1} \sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} V_{ik}^{-1} Y_{ik}. \end{aligned}$$

The partitioned scores $S_{k,\gamma}(\beta)$, $S_{k,\beta}(\beta)$ of $S_k(\theta)$ under the null and variance I_k of $S_{k,\gamma}(\hat{\beta}_k)$ are similarly defined as

$$S_{k,\gamma}(\beta) = \sum_{i=1}^{n_k} Z_{ik}^{\mathsf{T}} V_{ik}^{-1} (Y_{ik} - W_{ik}\beta),$$

$$S_{k,\beta}(\beta) = \sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} V_{ik}^{-1} (Y_{ik} - W_{ik}\beta)$$

and

$$\begin{aligned} I_k &= \Gamma_{k,\gamma\gamma} - \Gamma_{k,\gamma\beta}\Gamma_{k,\beta\beta}^{-1}\Gamma_{k,\beta\gamma} \\ &= \sum_{i=1}^{n_k} Z_{ik}^{\mathsf{T}} V_{ik}^{-1} Z_{ik} - \left(\sum_{i=1}^{n_k} Z_{ik}^{\mathsf{T}} V_{ik}^{-1} W_{ik} \right) \left(\sum_{i=1}^{n_k} W_{ik}^{\mathsf{T}} V_{ik}^{-1} W_{ik} \right)^{-1} \left(\sum_{i=1}^{n_k} Z_{ik}^{\mathsf{T}} V_{ik}^{-1} W_{ik} \right)^{\mathsf{T}} \end{aligned} \tag{4.13}$$

where Γ_k is defined and partitioned the same as (4.10).

Following the arguments developed by Lee, Kim and Tsiatis (1996), we can establish the joint distribution of sequentially computed score statistics $n^{-1/2}S_{k,\gamma}(\hat{\beta}_k)$, $k = 1,\ldots,K$, and can show an independent increments structure. When the underlying distribution is normal, the joint distribution of $S_{k,\gamma}(\hat{\beta}_k)$, $k = 1,\ldots,K$, is multivariate normal with mean μ and

$$\mathrm{Cov}\{S_{k,\gamma}(\hat{\beta}_k), S_{l,\beta}(\hat{\beta}_l)\} = \mathrm{Var}\{S_{k,\gamma}(\hat{\beta}_k)\} = I_k,\ 1 \le k \le l \le K$$

where μ is 0 under H_0 and $\delta\bar{I}_k$ under local alternatives. Furthermore, under suitable regularity conditions for a non-normal underlying distribution, the asymptotic joint distribution of $n^{-1/2}S_{k,\gamma}(\hat{\beta}_k)$, $k = 1,\ldots,K$, is multivariate normal with mean μ and covariance matrix

$$\mathrm{Cov}_A\{n^{-1/2}S_{k,\gamma}(\hat{\beta}_k), n^{-1/2}S_{l,\beta}(\hat{\beta}_l)\} = \mathrm{Var}_A\{n^{-1/2}S_{k,\gamma}(\hat{\beta}_k)\} = \bar{I}_k,\ 1 \le k \le l \le K$$

where $\bar{I}_k = \lim_{n\to\infty} n^{-1}I_k$ and μ is 0 under H_0 and $\delta\bar{I}_k$ under local alternatives. When I_k or $\bar{I}_k$ is unknown, a consistent estimator can be obtained from (4.13) by substituting V_{ik} with $(Y_{ik} - X_{ik}\hat{\theta}_k)(Y_{ik} - X_{ik}\hat{\theta}_k)^\top$ where $\hat{\theta}_k$ is the generalized least squares estimator of θ.

A random effects model can be also applied to construct a sequential procedure to test the null hypothesis $H_0 : \gamma = 0$. Instead of the model (4.4), consider a random effects model

$$Y_{ik} = Z_{ik}\gamma + W_{ik}\beta_i + \epsilon_{ik},\ i = 1,\ldots,n_k \text{ and } k = 1,\ldots,K$$

where γ is a fixed effect parameter, $\epsilon_{ik} \sim N(0,\Sigma_{ik})$ and $\beta_i \sim N(\beta,\Sigma_\beta)$ are all independent. The parameter β_i can be interpreted as participant effects parameter. This model can also be written as $Y_{ik} = Z_{ik}\gamma + W_{ik}\beta + W_{ik}\beta_i^* + \epsilon_{ik}$ where β is also fixed and $\beta_i^* \sim N(0,\Sigma_\beta)$ so that it is included in the model (4.4) with $V_{ik} = W_{ik}\Sigma_\beta W_{ik}^\top + \Sigma_{ik}$.

4.2.2 Semiparametric models

In this subsubsection, we review the results of Lee et al. (1996). Assume that at the k^{th} interim analysis, the marginal mean of Y_{ik} given X_{ik} is

$$E(Y_{ik}|X_{ik}) = \mu_{ik}(\theta) = g(X_{ik},\theta)$$

where g is a known function. Denote a working variance to be used instead of the unknown true variance $V_{ik} = \mathrm{Var}(Y_{ik}|X_{ik})$ by $v_{ik}(\theta,\alpha)$ with additional variance parameters α. Then, the score vector or generalized estimating equations has the form

$$S_k(\theta,\alpha) = \sum_{i=1}^{n_k} S_{ik}(\theta,\alpha) = \sum_{i=1}^{n_k} D_{ik}(\theta)^\top v_{ik}^{-1}(\theta,\alpha)(Y_{ik} - \mu_{ik}(\theta))$$

where $D_{ik}(\theta) = \partial \mu_{ik}(\theta)/\partial\theta$. Note that this score vector is reduced to (4.12) when $\mu_{ik}(\theta) = X_{ik}\theta$ and $v_{ik}(\theta,\alpha) = V_{ik}$, and hence, it can be regarded as a generalization of the least squares methods in Subsection 4.1.

When a consistent estimator $\hat{\alpha}$ of α is available, Liang and Zeger (1986) showed that $S_k(\theta,\hat{\alpha})$ is asymptotically equivalent to $S_k(\theta,\alpha)$, and hence the asymptotic properties regarding the inference on θ remain unchanged when using $S_k(\theta,\hat{\alpha})$ instead of $S_k(\theta,\alpha)$. We assume that α is known or a consistent estimator $\hat{\alpha}$ is available, and denote generalized estimating equations estimators of θ for both cases as the same $\hat{\theta}_k$. Note that $\hat{\theta}_k$ is consistent. For more details about estimation of α, refer to, for example, Crowder (1995) and Lee et al. (1996).

Partition $S_k(\theta,\alpha)$ as $\{S_{k,\gamma}(\gamma,\beta,\alpha), S_{k,\beta}(\gamma,\beta,\alpha)^{\intercal}\}^{\intercal}$ and let $\hat{\beta}_k$ be the restricted generalized estimating equations estimator of β under the null hypothesis. Then, as shown by Rotnitzky and Jewell (1990), it can be shown that the score statistic $S_{k,\gamma}(0,\hat{\beta}_k,\alpha)$ is asymptotically equivalent to $T_k(0,\beta,\alpha)$ where

$$T_k(\gamma,\beta,\alpha) = S_{k,\gamma}(\gamma,\beta,\alpha) - \Gamma_{k,\gamma\beta}\Gamma_{k,\beta\beta}^{-1} S_{k,\beta}(\gamma,\beta,\alpha) \tag{4.14}$$

and $\Gamma_{k,\gamma\beta}$ and $\Gamma_{k,\beta\beta}$ are submatrices of the partitioned matrix of Γ_k,

$$\Gamma_k = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n_k} D_{ik}(\theta)^{\intercal} v_{ik}^{-1}(\theta,\alpha) D_{ik}(\theta) = \begin{bmatrix} \Gamma_{k,\gamma\gamma} & \Gamma_{k,\gamma\beta} \\ \Gamma_{k,\beta\gamma} & \Gamma_{k,\beta\beta} \end{bmatrix}.$$

That is,

$$S_{k,\gamma}(0,\hat{\beta}_k,\alpha) \simeq S_{k,\gamma}(0,\beta,\alpha) - \Gamma_{0k,\gamma\beta}\Gamma_{0k,\beta\beta}^{-1} S_{k,\beta}(0,\beta,\alpha) = T_k(0,\beta,\alpha) \tag{4.15}$$

where the subscript 0 means "evaluated at $\gamma = 0$", or equivalently, "evaluated under the null hypothesis".

Since the score vector $n^{-1/2}S_k(\theta,\alpha)$ has a form of sum of independent variables, the asymptotic distribution of $n^{-1/2}S_k(\theta\alpha)$ is multivariate normal with mean 0 and variance

$$\Omega_k = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n_k} D_{ik}(\theta)^{\intercal} v_{ik}^{-1}(\theta,\alpha) V_{ik}\, v_{ik}^{-1}(\theta,\alpha) D_{ik}(\theta).$$

By asymptotic normality of $n^{-1/2}S_k(\theta,\alpha)$ together with the linear equation (4.14), the asymptotic joint distribution of $\{n^{-1/2}T_k(\gamma,\beta,\alpha), k = 1,\ldots,K\}$ becomes multivariate normal with mean 0. The asymptotic covariance of $n^{-1/2}T_k(\gamma,\beta,\alpha)$ and $n^{-1/2}T_l(\gamma,\beta,\alpha)$, for $k \le l$, is given by

$$M_{kl} = \Omega_{kl,\gamma\gamma} + \Gamma_{k,\gamma\beta}\Gamma_{k,\beta\beta}^{-1}\Omega_{kl,\beta\beta}\Gamma_{l,\beta\beta}^{-1}\Gamma_{l,\beta\gamma} - \Omega_{kl,\gamma\beta}\Gamma_{l,\beta\beta}^{-1}\Gamma_{l,\beta\gamma} - \Gamma_{k,\gamma\beta}\Gamma_{k,\beta\beta}^{-1}\Omega_{kl,\beta\gamma}$$

where $\Omega_{kl,\gamma\gamma}$, $\Omega_{kl,\gamma\beta}$, $\Omega_{kl,\beta\gamma}$ and $\Omega_{kl,\beta\beta}$ are submatrices of the partitioned matrix

$$\Omega_{kl} = \begin{bmatrix} \Omega_{kl,\gamma\gamma} & \Omega_{kl,\gamma\beta} \\ \Omega_{kl,\beta\gamma} & \Omega_{kl,\beta\beta} \end{bmatrix}$$

and

$$\Omega_{kl} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n_k} D_{ik}(\theta)^{\mathsf{T}} v_{ik}^{-1}(\theta,\alpha) V_{ikl}\, v_{il}^{-1}(\theta,\alpha) D_{il}(\theta).$$

Note that V_{ikl} denotes the true covariance matrix of Y_{ik} and Y_{il}. When the true variance functions are correctly specified, as shown in Lee et al. (1996), the asymptotic covariances M_{kl} $(1 \le k \le l \le K)$ are reduced to I_k,

$$I_k = \Gamma_{k,\gamma\gamma} - \Gamma_{k,\gamma\beta}\Gamma_{k,\beta\beta}^{-1}\Gamma_{k,\beta\gamma} = \mathrm{Var}_A\{n^{-1/2}T_k(\gamma,\beta,\alpha)\},$$

indicating an asymptotic independent increments structure.

By applying similar arguments to the equation (4.15), it can be shown that the asymptotic joint distribution of sequential score statistics $n^{-1}S_{k,\gamma}(0,\hat{\beta}_k,\alpha)$, $k = 1,\ldots,K$, is multivariate normal with mean 0 and covariance M_{0kl}, $k,l = 1,\ldots,K$. Furthermore, with a correct specification of the variance functions, we have $M_{0kl} = I_{0k} = \mathrm{Var}_A\{n^{-1/2}S_{k,\gamma}(0,\hat{\beta}_k,\alpha)\}$, which establish an asymptotic independent increments structure of sequentially computed score statistics.

The asymptotic variances Γ_k and Ω_k can be estimated consistently by evaluating $D_{ik}(\theta)$ and $v_{ik}(\theta,\alpha)$ at the consistent estimators $\hat{\alpha}$ and $\hat{\theta}_k$ and by substituting $\{Y_{ik} - \mu_{ik}(\hat{\theta}_k)\}\{Y_{ik} - \mu_{ik}(\hat{\theta}_k)\}^{\mathsf{T}}$ for V_{ik}. Under the null hypothesis, we use $\hat{\theta}_k = (0,\hat{\beta}_k^{\mathsf{T}})^{\mathsf{T}}$. As pointed out by Lee et al. (1996), these consistent estimators also lead to an asymptotic independent increments structure of sequentially computed $n^{-1/2}T_k(\gamma,\beta,\alpha)$ and $n^{-1}S_{k,\gamma}(0,\hat{\beta}_k,\alpha)$ when the variance functions are correctly specified.

4.3 Failure time data

In this subsection, we review the results for a general parametric model, Cox proportional hazards model by Cox (1972), and accelerated failure time model of Lin (1992), in the framework of counting process and martingale integration which can be referred to, for example, Fleming and Harrington (1991) and Anderson et al. (1993)

First, consider the notations for failure time data. Assume that n patients enter the trial at times $e_1,\ldots,e_n$ which are considered as constants. Each patient i has a potential failure time T_i, potential censoring time C_i, treatment indicator Z_i and covariate vector $W_i = (W_{i1},\cdots,W_{ip})$. It is assumed that T_i and C_i are conditionally independent given $\{Z_i,W_i\}$ and $\{T_i,C_i,Z_i,W_i\}$, $i = 1,\ldots,n$, are identically and independently distributed. If the data were analyzed at time t, the observable random variables would be $\{X_i(t),\Delta_i(t),Z_i,W_i\}$ for all $i = 1,\ldots,n$ such that $e_i \le t$. Here $X_i(t) = \min(T_i,C_i,t-e_i)$ is the time to failure or censoring, and $\Delta_i(t) = I\{T_i < \min(C_i,t-e_i)\}$ denotes the failure

indicator. For simplicity, we assume that the covariate vector W_i is time-invariant, but the same results are obtained for a time-varying covariate, as shown in Gu and Ying (1995) for the proportional hazards model.

We assume a hazard function $\lambda(u,Z_i,W_i,\theta)$ with $\theta=(\gamma,\beta^{\mathsf{T}})^{\mathsf{T}}$ where γ is a treatment effect parameter and β is a vector of nuisance parameters denoting covariate effects. As in the previous sections, we are interested in testing the null hypotheses (4.1). For notational simplicity, we set $R_i=(Z_i,W_i^{\mathsf{T}})^{\mathsf{T}}$.

It is convenient to express the failure time data in terms of counting process notation. Define the counting process of observed death for the ith patient at analysis time t by $N_i(u,t)=I\{X_i(t)\le u,\Delta_i(t)=1\}$ for $u\ge 0$. Note that whenever $e_i>t$, $N_i(u,t)=0$ for $u\ge 0$ and when $e_i\le t$, $N_i(u,t)=N_i(t,t)$ for $u\ge t$. Similarly, the at-risk process $Y_i(u,t)=I\{X_i(t)\ge u\}$, which is the indicator of whether the ith patient is at risk u units after entry into the study if the data were analyzed at calendar time t.

With the filtration $\mathcal{F}(u)$, $u\ge 0$, defined by Tsiatis et al. (1995), denote the $\mathcal{F}(u)$ martingale process associated with $N_i(u,t)$ by

$$M_i(u,t)=N_i(u,t)-\int_0^u \lambda(x,R_i,\theta)Y_i(x,t)dx.$$

Also, following the same procedure as in Tsiatis et al. (1995), we define the counting process of death observed between two successive analysis times t_k and t_{k-1} as $DN_i(u,t_1)=N_i(u,t_1)$ and $DN_i(u,t_k)=N_i(u,t_k)-N_i(u,t_{k-1})$, $k=2,\ldots,K$, where $t_1<\cdots<t_K$ denote the analysis times. Let $DY_i(u,t_k)=Y_i(u,t_k)-Y_i(u,t_{k-1})$, then the martingale process associated with $DN_i(u,t)$ can be written by

$$DM_i(u,t)=DN_i(u,t)-\int_0^u \lambda(x,R_i,\theta)DY_i(x,t)dx.$$

Note that since any two processes of $DN_i(u,t_k)$, $k=1,\ldots,K$, will not take jumps at the same time, $DM_i(u,t_k)$ and $DM_i(u,t_l)$ are orthogonal, that is, $\mathrm{Cov}\{DM_i(u,t_k),DM_i(u,t_l)\}=0$ if $k\ne l$. Also, note that $M_i(u,t_k)=\sum_{j=1}^k DM_i(u,t_j)$ for $k=1,\ldots,K$. So far, we defined the data structure, and related counting processes and martingales. We will use this common notations in the next subsubsections.

4.3.1 Parametric regression models

Assume that the hazard function $\lambda(u,R_i,\theta)$ is known, then using standard results for failure time data, the likelihood of the data available at time t is proportional to

$$L(t,\theta)=\Pi_{(i:e_i<t)}[\lambda\{X_i(t),R_i,\theta\}]^{\Delta_i(t)}\exp\left\{-\int_0^{X_i(t)}\lambda(u,R_i,\theta)du\right\}.$$

With the counting process notations, we can express the score vector at analysis time t as

$$\begin{aligned} S(t,\theta) &= \sum_{i=1}^{n} \int_0^{\infty} h(u,R_i,\theta)\{dN_i(u,t) - d\mu_i(u,t,\theta)\} \\ &= \sum_{i=1}^{n} \int_0^{\infty} h(u,R_i,\theta) dM_i(u,t) \end{aligned} \tag{4.16}$$

or equivalently, at analysis time t_k,

$$\begin{aligned} S(t_k,\theta) &= \sum_{j=1}^{k} \sum_{i=1}^{n} \int_0^{\infty} h(u,R_i,\theta)\{dDN_i(u,t_j) - dD\mu_i(u,t,\theta)\} \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} \int_0^{\infty} h(u,R_i,\theta) dDM_i(u,t_j) \end{aligned} \tag{4.17}$$

where

$$h(u,R_i,\theta) = \partial \log \lambda(u,R_i,\theta)/\partial\theta,$$
$$d\mu_i(u,t,\theta) = \lambda(u,R_i,\theta) Y_i(u,t) du$$

and

$$dD\mu_i(u,t,\theta) = \lambda(u,R_i,\theta) DY_i(u,t) du.$$

Denote the part taken from the second sum in (4.17) by S_j. Then, by the standard arguments for counting processes, e.g. in Fleming and Harrington (1991), the vector of martingale integrals S_j is also martingale, and we have

$$E\{dN_i(u,t)|R_i\} = d\mu_i(u,t,\theta)$$

and

$$E\{dDN_i(u,t)|R_i\} = dD\mu_i(u,t,\theta)$$

so that

$$E\{dM_i(u,t)|R_i\} = E\{dDM_i(u,t)|R_i\} = 0,$$
$$\mathrm{Var}\{dN_i(u,t)|R_i\} = \mathrm{Var}\{dM_i(u,t)|R_i\} = d\mu_i(u,t,\theta),$$

and

$$\mathrm{Var}\{dDN_i(u,t)|R_i\} = \mathrm{Var}\{dDM_i(u,t)|R_i\} = dD\mu_i(u,t,\theta).$$

Hence, S_j has mean 0 and variance

$$\mathrm{Var}(S_j) = \sum_{i=1}^{n} \int_0^{\infty} h(u,R_i,\theta) h(u,R_i,\theta)^{\mathsf{T}} dD\mu_i(u,t_j,\theta).$$

Furthermore, since $\mathrm{Cov}\{DM_i(u,t_k), DM_i(u,t_l)\} = 0$ if $k \neq l$ and observations are independent, S_j, $j = 1,\ldots,k$, are uncorrelated, and hence the score vector $S(t_k,\theta)$ has a similar form to (4.2), sum of uncorrelated variables. Applying the martingale central limit theorem, we can show that the asymptotic joint distribution of $n^{-1/2}S(t_k,\theta)$, $k = 1,\ldots,K$, is multivariate normal with mean 0. It can also be shown that

$$\mathrm{Cov}_A\{n^{-1/2}S(t_k,\theta), n^{-1/2}S(t_l,\theta)\} = \mathrm{Var}_A\{n^{-1/2}S(t_k,\theta)\} = \Gamma(t_k),\ 1 \leq k \leq l \leq K,$$

which indicates an independent increments structure. Here the asymptotic variance

$$\Gamma(t_k) = \lim_{n\to\infty} \int_0^\infty n^{-1} \sum_{i=1}^{n} h(u,R_i,\theta)h(u,R_i,\theta)^{\tau} d\mu_i(u,t_k,\theta). \tag{4.18}$$

Now, partition the score vector $S(t,\theta)$ as $\{S_\gamma(t,\gamma,\beta), S_\beta(t,\gamma,\beta)^{\tau}\}^{\tau}$, where $S_\gamma(t,\gamma,\beta) = \partial \log L(t,\gamma,\beta)/\partial\gamma$ and $S_\beta(t,\gamma,\beta) = \partial \log L(t,\gamma,\beta)/\partial\beta$. Then the score test of the null hypothesis $H_0 : \gamma = 0$ in the presence of nuisance parameters β, evaluated at calendar time t, is given by $S_\gamma(t,0,\hat{\beta}_t)$ where $\hat{\beta}_t$ is the restricted maximum likelihood estimator of β when $\gamma = 0$. Using standard results of likelihood theory, Cox and Hinkley (1974, Sec 9.3), the score test $S_\gamma(t,0,\hat{\beta}_t)$ is asymptotically equivalent to

$$T(t,0,\beta) = S_\gamma(t,0,\beta) - \Gamma_{\gamma\beta}(t)\Gamma_{\beta\beta}^{-1}(t)S_\beta(t,0,\beta)$$

where $\Gamma_{\gamma\beta}$ and $\Gamma_{\beta\beta}$ are submatrices of the partitioned matrix of $\Gamma_0(t)$, which is $\Gamma(t)$ in (4.18) evaluated at $\gamma = 0$,

$$\Gamma_0(t) = \begin{bmatrix} \Gamma_{\gamma\gamma}(t) & \Gamma_{\gamma\beta}(t) \\ \Gamma_{\beta\gamma}(t) & \Gamma_{\beta\beta}(t) \end{bmatrix}.$$

Since $n^{-1/2}T(t,0,\beta)$ is a linear combination of the elements of the score vector $n^{-1/2}$ $S(t,\theta)$, which converges in distribution to a multivariate normal with independent increments, this implies that $n^{-1/2}T(t,0,\beta)$ also converges in distribution to a normal with mean μ and variance $I(t)$, where $I(t) = \Gamma_{\gamma\gamma}(t) - \Gamma_{\gamma\beta}(t)\Gamma_{\beta\beta}^{-1}(t)\Gamma_{\beta\gamma}(t)$, and $\mu = 0$ under the null hypothesis and $\mu = \delta I(t)$ under the local alternatives defined in Subsubsection 4.2.1. Therefore, following the same arguments as in the previous sections, we can show that the asymptotic joint distribution of $\{n^{-1/2}S_\gamma(t_k,0,\hat{\beta}_{t_k}), k = 1,\ldots,K\}$ is multivariate normal with mean μ and covariance $(1 \leq k \leq l \leq K)$

$$\mathrm{Cov}_A\{n^{-1/2}S_\gamma(t_k,0,\hat{\beta}_{t_k}), n^{-1/2}S_\gamma(t_l,0,\hat{\beta}_{t_l})\} = \mathrm{Var}_A\{n^{-1/2}S_\gamma(t_k,0,\hat{\beta}_{t_k})\} = I(t_k), \tag{4.19}$$

which implies an independent increments structure of the asymptotic joint distribution.

Note that $h(u,R_i,\theta)$ does not depend on the calendar time t, and this make it much easier to establish the independent increments structure in (4.18) and (4.19). In fact, it can be shown that when we use a weighted score vector with a weight function

$Q(u,t,\theta)$ which converges in probability to a function $q(u,t,\theta)$, the independent increments structure holds as long as $q(u,t,\theta)$ does not depend on the calendar time t. Therefore, choosing a suitable weight function, we can construct a group sequential test having asymptotic normality and independent increments structure. This may be particularly useful when the efficient test is difficult to be built explicitly, as found in the accelerated failure time model. For weighted tests, the limiting optimal weight function is proportional to the limit of $h(u,R_i,\theta) = \partial \log \lambda(u,R_i,\theta)/\partial\theta$ because the score vectors in (4.16) and (4.17) are efficient scores. Tsiatis (1982) and Lin (1992) also showed, for the proportional hazards model and accelerated failure time model, that the limiting weight functions preserve the independent increments structure.

It is also interesting to note that the score vectors given by (4.16) and (4.17) can be regarded as the score vectors based on the generalized estimating equations accommodating time dependent structure of the failure time data by a stochastic integral. Considering $dN_i(u,t)$ as the ith observation and expressing $h(u,R_i,\theta)$ in (4.16) as

$$h(u,R_i,\theta) = \{\partial d\mu_i(u,t,\theta)/\partial\theta\}/\mathrm{Var}\{dN_i(u,t)|R_i\},$$

we have the generalized estimating equations

$$S(t,\theta) = \sum_{i=1}^{n} \int_0^\infty n^{-1}\{\partial d\mu_i(u,t,\theta)/\partial\theta\}\left[\mathrm{Var}\{dN_i(u,t)|R_i\}\right]^{-1}\{dN_i(u,t) - d\mu_i(u,t,\theta)\}.$$

In this framework, choosing a weight function corresponds to choosing a working variance.

4.3.2 Proportional hazards models

Consider the Cox proportional hazards model where the hazard function $\lambda(u,R_i,\theta)$ is given by

$$\lambda(u,R_i,\theta) = \lambda_0(u)\exp(\theta^{\mathrm{T}}R_i)$$

where λ_0 is an arbitrary baseline hazard fuction. We can express the score vector based on the partial likelihood (Cox, 1975) at the analysis time t as

$$\begin{aligned} U(t,\theta) &= \sum_{i=1}^{n} \int_0^\infty \{R_i - \bar{R}(u,t,\theta)\} dN_i(u,t) \\ &= \sum_{i=1}^{n} \int_0^\infty \{R_i - \bar{R}(u,t,\theta)\} dM_i(u,t) \end{aligned} \tag{4.20}$$

where $\bar{R}(u,t,\theta) = \sum_{i=1}^n R_i Y_i(u,t)\exp(\theta^{\mathrm{T}}R_i)/\sum_{i=1}^n Y_i(u,t)\exp(\theta^{\mathrm{T}}R_i)$.

The partial likelihood score vector (4.20) has the same form as the maximum likelihood score vector (4.16) if $h(u,R_i,\theta)$ in (4.16) is replaced with $R_i - \bar{R}(u,t,\theta)$. Though $R_i - \bar{R}(u,t,\theta)$ may depend on the calendar time t, as shown in Jennison and Turnbull

(1997), the independent increments structure of the score vector $U(t,\theta)$ still holds. Therefore, the arguments in Subsubsection 4.3.1 can be applied to produce the following results:

Denote the score test statistic for the null hypothesis by $U_\gamma(t,0,\hat{\beta})$ where $U_\gamma(t,\gamma,\beta)$ is the first element of the partitioned score vector $U(t,\theta) = \{U_\gamma(t,\gamma,\beta), U_\beta(t,\gamma,\beta)^\top\}^\top$ and $\hat{\beta}$ is the restricted maximum partial likelihood estimator of β when $\gamma = 0$. Then the asymptotic joint distribution of $\{n^{-1/2}U_\gamma(t_k,0,\hat{\beta}_{t_k}), k = 1,\ldots,K\}$ is multivariate normal with mean μ and covariance $(1 \le k \le l \le K)$

$$\mathrm{Cov}_A\{n^{-1/2}U_\gamma(t_k,0,\hat{\beta}_{t_k}), n^{-1/2}U_\gamma(t_l,0,\hat{\beta}_{t_l})\} = \mathrm{Var}_A\{n^{-1/2}U_\gamma(t_k,0,\hat{\beta}_{t_k})\} = I(t_k),$$

which implies an independent increments structure of the asymptotic joint distribution. Here, $I(t) = \Gamma_{\gamma\gamma}(t) - \Gamma_{\gamma\beta}(t)\Gamma^{-1}_{\beta\beta}(t)\Gamma_{\beta\gamma}(t)$ and

$$\Gamma_0(t) = \begin{bmatrix} \Gamma_{\gamma\gamma}(t) & \Gamma_{\gamma\beta}(t) \\ \Gamma_{\beta\gamma}(t) & \Gamma_{\beta\beta}(t) \end{bmatrix},$$

which is obtained by evaluating, at $\gamma = 0$, $\Gamma(t)$,

$$\Gamma(t) = \lim_{n\to\infty} \int_0^\infty n^{-1} \sum_{i=1}^n \{R_i - \bar{R}(u,t,\theta)\}\{R_i - \bar{R}(u,t,\theta)\}^\top Y_i(u,t)\lambda_0(u)\exp(\theta^\top R_i)du.$$

The variance matrix $I(t)$ can be consistently estimated by substituting β and $\lambda_0(u)$ in $\Gamma_0(t)$ with $\hat{\beta}_t$ and the Breslow estimator evaluated under the null hypothesis,

$$\hat{\lambda}_0(u,0,\hat{\beta}_t) = \sum_{i=1}^n dN_i(u,t) \Big/ \sum_{i=1}^n Y_i(u,t)\exp(\hat{\beta}_t^\top W_i).$$

In general, the Breslow estimator is given by

$$\hat{\lambda}_0(u,\hat{\theta}) = \sum_{i=1}^n dN_i(u,t) \Big/ \sum_{i=1}^n Y_i(u,t)\exp(\hat{\theta}^\top R_i)$$

where $\hat{\theta}$ is the maximum partial likelihood estimator of θ, and if there are no covariates, it becomes the Nelson-Aalen estimator by Aalen (1978).

As mentioned in Subsection 4.1, we can consider the weighted score vector $U_Q(t,\theta)$ with a weight function $Q(u,t,\theta)$,

$$U_Q(t,\theta) = \sum_{i=1}^n \int_0^\infty Q(u,t,\theta)\{R_i - \bar{R}(u,t,\theta)\}dN_i(u,t),$$

and we can show that the independent increments structure of sequentially computed score statistics holds when $Q(u,t,\theta)$ converges in probability to a limit $q(u,\theta)$ free of t.

When there are no covariates, the weighted score vector leads to the well known two-sample weighted logrank tests which were studied by Tsiatis (1982).

For a given θ, let $\hat{\lambda}_0(u,\theta)$ denote the Breslow estimator. Then, comparing the partial likelihood score vector (4.20) with the maximum likelihood score vector (4.16), we can show that $U(t,\theta) = S(t,\theta,\hat{\lambda}_0)$ where $S(t,\theta,\hat{\lambda}_0)$ is the score vector obtained by replacing $\lambda_0(u)$ with $\hat{\lambda}_0(u,\theta)$ in (4.16) or (4.17). It seems that the score vector (4.16) can also be expressed in the generalized estimating equations framework as

$$U(t,\theta) = \sum_{i=1}^{n} \int_0^\infty \{\partial d\hat{\mu}_i(u,t,\theta)/\partial\theta\} \left[\widehat{\mathrm{Var}}\{dN_i(u,t)|R_i\}\right]^{-1} \{dN_i(u,t) - d\mu_i(u,t,\theta)\}$$

where $d\hat{\mu}_i(u,t,\theta) = \widehat{\mathrm{Var}}\{dN_i(u,t)|R_i\} = Y_i(u,t)\hat{\lambda}_0(u,\theta)\exp(\theta^{\mathsf{T}}R_i)$. By the consistency of the Breslow estimator $\hat{\lambda}_0(u,\theta)$, this score vector can be regarded as the generalized estimating equations score vector obtained when the true variances are consistently estimated.

4.3.3 Accelerated failure time models

Consider the linear model

$$T_i = \theta^{\mathsf{T}}R_i + \epsilon_i,\ i = 1,\ldots,n$$

where ϵ_i are independent with a common hazard function λ_0. Here, T_is are usually log transformed observation of the original nonnegative failure time data so that they are allowed to have negative values. Further, assume that the treatment indicator Z_i is independent of the covariates W_i as in usual clinical trials.

For a given λ_0, the efficient score vector (4.16) can be written as

$$S(t,\theta) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} R_i\{\lambda_0'(u-\theta^{\mathsf{T}}R_i)/\lambda_0(u-\theta^{\mathsf{T}}R_i)\}\{dN_i(u,t) - Y_i(u,t)\lambda_0(u-\theta^{\mathsf{T}}R_i)du\}.$$

When λ_0 is unknown, replacing $\lambda_0(u)$ with $\hat{\lambda}_0(u,t,\theta)$ and $\lambda_0'(u)/\lambda_0(u)$ with a weight function $Q(u,\theta)$, we have

$$S(t,\theta,\hat{\lambda}_0) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} Q(u,\theta)R_i\{dN_i(u+\theta^{\mathsf{T}}R_i,t) - Y_i(u+\theta^{\mathsf{T}}R_i,t)\hat{\lambda}_0(u,\theta)du\},$$

where $\hat{\lambda}_0(u,t,\theta)du = \sum_{i=1}^n dN_i(u+\theta^{\mathsf{T}}R_i,t)/\sum_{i=1}^n Y_i(u+\theta^{\mathsf{T}}R_i,t)$ is a Nelson-Aalen type estimator of $\lambda_0(u)$ at the analysis time t. Furthermore, let

$$\bar{R}(u,t,\theta) = \sum_{i=1}^{n} R_iY_i(u+\theta^{\mathsf{T}}R_i,t)/\sum_{i=1}^{n} Y_i(u+\theta^{\mathsf{T}}R_i,t).$$

Then $S(t,\theta,\hat{\lambda}_0)$ is equivalent to a rank score vector

$$\begin{aligned} U(t,\theta) &= \sum_{i=1}^{n}\int_{-\infty}^{\infty} Q(u,\theta)\{R_i-\bar{R}(u,t,\theta)\}dN_i(u+\theta^{\mathsf{T}}R_i,t) \\ &= \sum_{i=1}^{n}\int_{-\infty}^{\infty} Q(u,\theta)\{R_i-\bar{R}(u,t,\theta)\}dM_i(u+\theta^{\mathsf{T}}R_i,t) \end{aligned} \tag{4.21}$$

where $M_i(u+\theta^{\mathsf{T}}R_i,t) = N_i(u+\theta^{\mathsf{T}}R_i,t) - \int_{-\infty}^{u} Y_i(x+\theta^{\mathsf{T}}R_i,t)\lambda_0(x)dx$ is a martingale associated with the counting process $N_i(u+\theta^{\mathsf{T}}R_i,t)$ of the residual $X_i(t)-\theta^{\mathsf{T}}R_i$. This class of linear rank tests (4.21) were studied by Tsiatis (1990), Ritov (1990), Wei, Ying and Lin (1990b), and Lin (1992). As shown in Tsiatis (1990), note that the limiting optimal weight function is proportional to $\lambda_0'(u)/\lambda_0(u)$. This rank score vector can also be interpreted in the generalized estimating equations framework, as described in Subsubsection 4.3.2.

At a glance, it seems that the rank score (4.21) has the same form as those of efficient scores for the parametric model and the proportional hazards model, and hence that the same arguments as discussed in the previous sections can be applied. However, as pointed out in several researches such as Tsiatis (1990) and Lin, Wei and Ying (1998), because the rank score is a step function of θ, any exact solution of $U(t,\hat{\theta})=0$ may not exist. Therefore, $\hat{\theta}$ is defined as a value θ for which $U(t,\theta)$ changes sign or as a minimizer of $\| U(t,\theta) \|$ where $\| a \| = (a^{\mathsf{T}}a)^{1/2}$. For more discussions on this minimization problem, refer to Wei et al. (1990b) and Lin et al. (1998).

For simplicity, assume $Q(u,\theta)=1$ temporarily and let $E_i(t,\beta) = X_i(t)-\beta^{\mathsf{T}}W_i$ for $i=1,\ldots,n$. Further, define $N_i^*(u,t) = \Delta_i(t)I\{E_i(t,\beta)\le u\}$ and $Y_i^*(u,t) = I\{E_i(t,\beta)\ge u\}$. Then, under the null hypothesis, $U(t,\theta)$ is partitioned as

$$\begin{aligned} U_\gamma(t,\beta) &= \sum_{i=1}^{n}\int_{-\infty}^{\infty}\{Z_i-\bar{Z}(u,t,\beta)\}dN_i^*(u,t) \\ &= \sum_{i=1}^{n}\int_{-\infty}^{\infty}\{Z_i-\bar{Z}(u,t,\beta)\}dM_i^*(u,t) \end{aligned}$$

and

$$\begin{aligned} U_\beta(t,\beta) &= \sum_{i=1}^{n}\int_{-\infty}^{\infty}\{W_i-\bar{W}(u,t,\beta)\}dN_i^*(u,t) \\ &= \sum_{i=1}^{n}\int_{-\infty}^{\infty}\{W_i-\bar{W}(u,t,\beta)\}dM_i^*(u,t) \end{aligned}$$

where

$$\bar{Z}(u,t,\beta) = \sum_{i=1}^{n} Z_iY_i^*(u,t) \Big/ \sum_{i=1}^{n} Y_i^*(u,t),$$

$$\bar{W}(u,t,\beta) = \sum_{i=1}^{n} W_i Y_i^*(u,t) / \sum_{i=1}^{n} Y_i^*(u,t)$$

and

$$M_i^*(u,t) = N_i^*(u,t) - \int_{-\infty}^{u} Y_i^*(x,t)\lambda_0(x)dx,$$

which is a martingale. Note that the score function $U_\gamma(t,\beta)$ has the same form as the score functions $S_\gamma(t,0,\beta)$ and $U_\gamma(t,0,\beta)$ in Subsubsections 4.3.1 and 4.3.2, respectively, so that we can apply the similar arguments to establish the asymptotic results of $U_\gamma(t,\beta)$. That is, the asymptotic joint distribution of $\{n^{-1/2}U_\gamma(t_k,\beta), k = 1,\ldots,K\}$ is multivariate normal with mean 0 and covariance $(1 \le k \le l \le K)$

$$\mathrm{Cov}_A\{n^{-1/2}U_\gamma(t_k,\beta), n^{-1/2}U_\gamma(t_l,\beta)\} = \mathrm{Var}_A\{n^{-1/2}U_\gamma(t_k,\beta)\} = I(t_k),$$

where

$$I(t) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \{Z_i - \bar{Z}(u,t,\beta)\}^2 Y_i^*(u,t)\lambda_0(u)du. \tag{4.22}$$

Since the rank score $U(t,\theta)$ is a step function of θ, we can not apply the usual Taylor expansions to find out a test statistic asymptotically equivalent to $U_\gamma(t,\hat{\beta}_t)$, where the restricted estimator $\hat{\beta}_t$ is the minimizer of $\| U_\beta(t,\beta) \|$. Under the assumption of independence of Z_i and W_i, however, Lin (1992) showed that $U_\gamma(t,\hat{\beta}_t)$ is asymptotically equivalent to $U_\gamma(t,\beta)$. In this case, as shown in Lin (1992), we can simplify $I(t)$ in (4.22). Note that $\bar{Z}(u,t,\beta)$ converges in probability to $\mu_z = E(Z_i)$ and

$$E\left\{\int_{-\infty}^{\infty} Y_i^*(u,t)\lambda_0(u)du\right\} = E\{N_i(\infty,t)\} = \Pr\{\Delta_i(t) = 1\}.$$

Furthermore, Z_i are independent of the other variables so that $I(t) = \sigma_z^2 \Pr\{\Delta_i(t) = 1\}$, where $\sigma_z^2 = E\{(Z_i - \mu_z)^2\}$. Hence, we have that the asymptotic joint distribution of $\{n^{-1/2}U_\gamma(t_k,\hat{\beta}_{t_k}), k = 1,\ldots,K\}$ is multivariate normal with mean 0 and covariance matrix $\{\sigma^2(t_k,t_l); k,l = 1,\ldots,K\}$ where $\sigma^2(t,t') = \sigma^2(t,t) = \sigma_z^2 \Pr\{\Delta_i(t) = 1\}$ for $t \le t'$. Under the null hypothesis, denote $Q(u,\theta)$ and $\hat{\lambda}_0(u,t,\theta)$ by $Q(u,\beta)$ and $\hat{\lambda}_0(u,t,\beta)$. Then, for a given weight function $Q(u,\beta)$, the variance function $\sigma^2(t,t)$ can be consistently estimated by

$$\begin{aligned}
\hat{I}(t) &= n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\infty} Q^2(u,\hat{\beta}_t)\{Z_i - \bar{Z}(u,t,\hat{\beta}_t)\}^2 Y_i^*(u,t,\hat{\beta}_t)\hat{\lambda}_0(u,t,\hat{\beta}_t)du \\
&= n^{-1} \int_{-\infty}^{\infty} Q^2(u,\hat{\beta}_t) \left\{ \frac{\sum_{i=1}^n Z_j^2 Y_j^*(u,t,\hat{\beta}_t)}{\sum_{i=1}^n Y_j^*(u,t,\hat{\beta}_t)} - \bar{Z}^2(u,t,\hat{\beta}_t) \right\} \sum_{i=1}^{n} dN_i^*(u,t,\hat{\beta}_t)
\end{aligned}$$

where $Y_i^*(u,t,\hat{\beta}_t)$ and $N_i^*(u,t,\hat{\beta}_t)$ are obtained by substituting β with $\hat{\beta}_t$ in $Y_i^*(u,t)$ and $N_i^*(u,t)$, respectively.

5 Examples

5.1 Error spending based on information

As mentioned in Subsection 2.4, standard group sequential methods by Pocock (1977) and O'Brien and Fleming (1979) require equal increments of information at each interim analysis and a pre-specification of the maximum number of analyses. However, these conditions are often not met in practice. The error spending function approach of Lan and DeMets (1983) guarantees an overall type I error probability to a desired significance level without having to fix the number and times of repeated analyses in advance. When designing a study, the number and times of repeated analyses have to be fixed at least tentatively based on the projected duration of enrollment and follow-up and the desired frequency of interim analyses for possible early stopping. This is an issue of particular interest in designing clinical trials with failure time data. A natural approach is to use the notion of statistical information and design the trial as a maximum information trial as in Kim et al. (1995), Lee et al. (1996), and Scharfstein et al. (1997).

At the k^{th} interim analysis, $k = 1, \ldots, K$, denote the standardized score statistics by $S_k = S(\hat{\beta}_k)/\text{SE}\{S(\hat{\beta}_k)\}$ and the standardized Wald test by $W_k = \hat{\gamma}_k/\text{SE}(\hat{\gamma}_k)$ for testing the null hypothesis $\gamma = 0$ in the presence of the nuisance parameters β, where $S(\hat{\beta}_k)$ is the usual score statistics presented in the previous sections with the restricted estimator $\hat{\beta}_k$ of β under the null hypothesis. For the Wald test, $\hat{\gamma}_k$ is obtained from the estimating equations such as the maximum likelihood estimating equations, the least squares estimating equations, the generalized estimating equations and the rank type estimating equations described in the previous sections. Then the information I_k at the k^{th} interim analysis is defined by $I_{k,u} = \text{Var}\{S(\hat{\beta}_k)\}$ for the score test and $I_{k,e} = \{\text{Var}(\hat{\gamma}_k)\}^{-1}$ for Wald test. The information I_k can be estimated by replacing Var with $\widehat{\text{Var}}$. We denote it as $\hat{I}_k$.

For an error spending function $\alpha^\star(t)$ described in Subsection 2.4, we can use the information fraction t_k for the k^{th} interim analysis given by a ratio of the information at the k^{th} interim analysis to the maximum information I_K predetermined by design, i.e. $t_k = I_k/I_K$. At the time of the k^{th} interim analysis, I_k is obtained from the test statistics. The maximum information is defined as

$$I_K = \left(\frac{z_{\alpha/2} + z_\beta}{\gamma_A}\right)^2 \mathfrak{I}_F \tag{5.1}$$

where α and $1-\beta$ are the type I error and the power to be required, respectively; γ_A denotes the treatment effects under the alternative hypothesis; and $\mathfrak{I}_F$ is the so-called inflation factor. The required inflation in statistical information to compensate for the loss of power through multiple testing was discussed by Kim and DeMets (1987). The inflation factor $\mathfrak{I}_F$ is determined as a function of α, β and the number K and timing of repeated testing and depends on the selected error spending function or the group sequential method. Scharfstein et al. (1997) also provide a table of the inflation factors

for methods by Pocock (1977) and O'Brien and Fleming (1979) under various design schemes.

For a given I_K in (5.1), the critical value c_k is calculated by solving the equation

$$\Pr(|Z_1| \le c_1, \ldots, |Z_{k-1}| \le c_{k-1}, |Z_k| > c_k) = \alpha^\star(t_k) - \alpha^\star(t_{k-1}), \tag{5.2}$$

where $(Z_1, \ldots, Z_K)$ is multivariate normal with mean 0 and covariance

$$(I_k/I_l)^{1/2},\ 1 \le k \le l \le K. \tag{5.3}$$

Reboussin et al. (2000) provided programs for calculating group sequential boundaries using the Lan and DeMets (1983) method. The boundary values $b_{k,u}$ for the score test S_k and $b_{k,e}$ for the Wald test W_k are obtained by replacing t_k in (5.2) with $t_{k,u}$ and $t_{k,e}$, and replacing I_k/I_l in (5.3) with $\hat{I}_{k,u}/\hat{I}_{l,u}$ and $\hat{I}_{k,e}/\hat{I}_{l,e}$, respectively. Here $t_{k,u} = \hat{I}_{k,u}/I_K$ and $t_{k,e} = \hat{I}_{k,e}/I_K$. If $|S_k| > b_{k,u}$ for the score test and $|W_k| > b_{k,e}$ for the Wald test, one stops and rejects the null hypothesis. Note that the covariance (5.3) implies the independent increments structure so that we can use the recursion formula in Subsection 2.3.

5.2 Longitudinal data

To examine the finite sample properties of the "score" test and the Wald test for the semiparametric model for longitudinal data, we use a semiparametric model suggested by the data from the National Cooperative Gallstone Study (NCGS) in Schoenfield et al. (1981). For illustration, we consider only the comparison of cholesterol levels between the placebo (305 patients) group and the high-dose chenodiol (305 patients) group.

The four repeated cholesterol values are modeled as a linear function of the baseline cholesterol value (B_i) and the treatment indicator (T_i) for $i = 1, \ldots, n$ and $j = 1:4$ and the k^{th} interim analysis as

$$E(Y_{ijk}|X_{ik}) = \beta_1 T_i + \beta_{0j} I_{ijk} + \beta_{1j} I_{ijk} B_i.$$

The estimated covariance matrix of the score test statistics and the Wald test statistics over time are, respectively, as follows:

$$\begin{bmatrix} 0.1027 & 0.1044 & 0.1064 & 0.1075 \\ & 0.1490 & 0.1497 & 0.1495 \\ & & 0.1927 & 0.1824 \\ & & & 0.2217 \end{bmatrix}$$

and

$$\begin{bmatrix} 8.8741 & 6.2173 & 5.0184 & 4.3387 \\ & 6.1136 & 4.8675 & 4.1597 \\ & & 4.9627 & 4.0206 \\ & & & 4.1814 \end{bmatrix}.$$

These results confirm empirically the independent increments structure in the sequential test statistics as noted in (2.7) and (2.6) from (2.8), respectively.

5.3 Failure Time data

We describe a simulation study reported in Tsiatis et al. (1995) to illustrate how the group sequential tests for parametric model for failure time data work with moderate sample sizes that are typical in clinical trials. In the simulation, 100 patients were entered uniformly over a 10 year period, and each patient entering the trial in a staggered fashion was randomly allocated with equal probability to one of two treatments indicated by $Z = 0$ or 1. A failure time W_i for patient i was obtained as a function of treatment assignment Z_i and trial entry time E_i by generating an exponentially distributed random variable given by the exponential model with the hazard rate

$$\lambda(u|Z,E,\beta,\theta) = \exp(\theta_1 + \beta Z + \theta_2 E)$$

which is a function of both treatment and entry time. We considered a test of the null hypothesis of no treatment difference, $\beta = 0$, with the nuisance parameters $\theta_1 = 0$ and $\theta_2 = 0.1$.

We analyzed the accumulating data at four times after equal increments in calendar time, i.e. $t = 2.5, 5.0, 7.5$, and 10 years, using all the data available at those times. At each of the four times, we calculated the maximum likelihood estimate $\hat{\beta}(t)$, the score statistic $S_0\{t, \beta = 0, \hat{\theta}(\beta = 0)\}$, and the observed information $\{I^{00}(t)\}^{-1}$. The maximum information was set equal to the average $\{I^{00}(10)\}^{-1}$ obtained from 10,000 repetitions.

To empirically examine the type I error probability, we recorded the proportion of rejections for both the score test and the Wald test, using the Pocock and O'Brien-Fleming type error spending functions at the 0.05 level of significance in another set of 10,000 repetitions. With the group-sequential test based on the Pocock type error spending function $\alpha^{\star}_{\mathrm{P}}(t)$, 598 and 525 of the 10,000 simulations rejected the null hypothesis for the score test and Wald test, respectively. With the group-sequential test based on the O'Brien-Fleming type error spending function $\alpha^{\star}_{\mathrm{OF}}(t) = 2\{1 - \Phi(z_{\alpha/2}/\sqrt{t})\}$, 531 and 519 of the 10,000 simulations rejected the null hypothesis for the score test and the Wald test, respectively. These results seem to suggest that the Wald test produces the type I error probability close to the target significance level as compared to the score test. Also O'Brien-Fleming type group sequential test produces the type I error probability close to the target significance level as compared to the Pocock type group sequential test.

In order to verify the independent increments structure in the sequentially computed test statistics, we computed the empirical correlation matrix of the increments of the score test and the Wald test pre-multiplied by the observed information matrix. They are

$$\begin{bmatrix} 1 & -0.0169 & 0.0011 & 0.0002 \\ & 1 & -0.0059 & 0.0003 \\ & & 1 & -0.0176 \\ & & & 1 \end{bmatrix}$$

for the score test and

$$\begin{bmatrix} 1 & 0.0054 & -0.0028 & -0.0007 \\ & 1 & -0.0032 & -0.0025 \\ & & 1 & -0.0159 \\ & & & 1 \end{bmatrix}$$

for the Wald test. In both the score test and the Wald test, the simulation results appear to confirm the theory, as indicated by the off-diagonal entries being all very close to zero.

As a second example, we consider the Children's Cancer Group study 251 (CCG 251) in which 508 eligible children with untreated acute myeloid leukemia were enrolled between September 1979 and October 1983 in a staggered entry to receive an induction chemotherapy followed by either allogeneic bone marrow transplant or maintenance chemotherapy as reported in Lee and Sather (1995). Post-remission treatment was determined by whether patient had an HLA-matching sibling donor or not without randomization.

A total of 340 children achieved remission and were subsequently allocated to either transplant or chemotherapy. The primary outcome was disease-free survival from the end of induction chemotherapy. As there was apparent cure of disease in a substantial portion of children (30-45%), we analyzed disease-free survival using the mixture model with cure, also known as the cure rate model given by the survival function

$$S(t|Z,\beta,\theta) = \nu_Z + (1-\nu_Z)H_Z(t)$$

where the cure probability ν_Z is parametrized as

$$\nu_Z = \frac{\exp(\alpha+\beta Z)}{1+\exp(\alpha+\beta Z)}$$

and

$$H_Z(t) = \exp\{-\exp(\gamma_o+\gamma_1 Z)t\}^{\delta}.$$

Here β is the parameter of interest and $\theta = (\alpha,\gamma_0,\gamma_1,\delta)'$ is the nuisance parameter.

The study was originally conducted as a fixed sample trial. In order to illustrate the application of group sequential methods using the O'Brien-Fleming type error spending function $\alpha^{\star}_{\mathrm{OF}}(t) = 2\{1-\Phi(z_{\alpha/2}/\sqrt{t})\}$, we applied the score test and the Wald test at yearly intervals starting in October 1982 for three times with 255, 324, and 340 children. Tables ***1*** and ***2*** summarize the results for the score test and the Wald-type test, respectively.

Table 1: *Interim Analyses of CCG 251 with the Score Test.*

k	$\hat{\beta}$	$\widehat{\mathrm{Var}}(\hat{\beta})$	t_k	$\alpha^{\star}_{\mathrm{OF}}(t)$	$\lvert S(t_k)\rvert$	c_k	Reject H_0
1	2.77	4.63	0.307	0.0001	1.29	3.88	No
2	5.97	6.80	0.451	0.0017	2.29	3.14	No
3	9.50	13.38	0.888	0.0348	2.60	2.18	Yes

Table 2: *Interim Analyses of CCG 251 with the Wald Test.*

k	$\hat{\beta}$	$\widehat{\mathrm{Var}}(\hat{\beta})$	t_k	$\alpha^{\star}_{\mathrm{OF}}(t)$	$\lvert W(t_k)\rvert$	c_k	Reject H_0
1	0.515	0.177	0.374	0.0005	1.22	3.48	No
2	0.748	0.125	0.532	0.0043	2.12	2.87	No
3	0.684	0.071	0.931	0.0404	2.56	2.17	Yes

Unlike in Lee and Sather (1995) where the critical values had to be determined using multivariate normal integration as in Schervish (1984), the critical values in Tables ***1*** and ***2*** were determined using univariate normal integration thanks to independent increments. Note that the different test resulted in different estimates of the information fractions at each of the three interim analyses. With both the score test and the Wald test, the trial would have been terminated early in October 1984 with less than full information of 0.888 and 0.931, respectively.

6 Discussion

After the theoretical development in sequential analysis with the seminal work of Wald (1947), ethical imperatives of having to avoid unnecessary experimentation with human subjects in clinical trials motivated early pioneers such as Peter Armitage leading rapid development of sequential methods for clinical trials, e.g. including the first edition of the textbook "Sequential Medical Trials" by Peter Armitage in 1960. Soon there was a recognition, however, that classical sequential methods were not very realistic in most clinical trials and subsequently group sequential methods started to appear in the literature in 1970s.

In order for group sequential methods to be applied correctly, the joint distribution of sequential test statistics computed over time has to be known to determine the group sequential boundaries. In many settings the joint distribution turned out to be a multivariate normal distribution or asymptotically so. This required multivariate normal integration which can be challenging and applicable for up to seven dimensions. However, if the joint distribution has an independent increments structure in its covariance matrix, the multivariate integration reduces to univariate integration involving simple recursion in successive test statistics.

Many authors established the multivariate normality of the joint distributions of sequential test statistics. Many joint distributions turned out to have correlated increments between successive test statistics requiring multivariate normal integration. Examples include tests by Armitage et al. (1985), Geary (1988), Wei et al. (1990a), Lee and DeMets (1992), and Su and Lachin (1992) for longitudinal data and Gehan's test by Slud and Wei (1982) and logrank test under the accelerated failure model by Lin (1992) for failure time data.

Fortunately, joint distributions of many useful test statistics computed over time turn out to have independent increments, thus requiring only univariate integration based on convolution of two independent random variables. Independence of increments in the joint distribution of sequential test statistics was conjectured in Armitage (1975), but the theoretical development started with the initial work in Tsiatis (1981), followed by many noted in Section 3, and culminating with the most general results by Jennison and Turnbull (1997) and Scharfstein et al. (1997).

The limited simulation studies and the real clinical trials data analysis reported here show that the joint distributions of the sequential test statistics investigated have independent increments even for moderate sample sizes. This affirms that standard group sequential methods can be readily applied in interim analysis for possible early stopping of clinical trials in chronic diseases with the very common primary outcome of longitudinal data and failure time data.

Acknowledgements

We thank Professor Pere Puig Casado for inviting us to write this review paper. This is based partly on a presentation given at "A Symposium Honoring Professor Anastasios A. Tsiatis" held on 13 July 2013 in Raleigh, North Carolina, USA, to recognize his contributions in statistical theory and methods on the occasion of his 65th birthday.

Research reported by KMK and AAT was supported in part by grants from the National Institutes of Health at the time of publication.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6, 701–726.

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

Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, 10, 89–100.

Armitage, P. (1954). Sequential tests in prophylactic and therapeutic trials. *Quarterly Journal of Medicine*, 23(91), 255–274.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika*, 44, 9–26.

Armitage, P. (1975). *Sequential Medical Trials*, 2nd ed. Wiley, New York.

Armitage, P. (1990). Sequential methods. In Hinkley, D. V., Reid, N., and Snell, E. J. (eds), *Statistical Theory and Modeling: In Honor of Sir David Cox, FRS*. Chapman and Hall, London.

Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society*, Series A, 132, 235–244.

Armitage, P., Stratton, I. M., and Worthington, H. V. (1985). Repeated significance tests for clinical trials with a fixed number of patients and variable follow-up. *Biometrics*, 41, 353–359.

Bross, I. (1952). Sequential medical plans. *Biometrics*, 8, 188–205.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, Series B, 34, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62, 269–276.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Crowder, M. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika*, 82, 407–410.

DeMets, D. L. and Lan, K. K. G. (1994). Interim analysis: the alpha spending function approach. *Statistics in Medicine*, 13, 1341–1352.

Dodge, H. F. and Romig, H. G. (1929). A method of sampling inspection. *Bell System Technical Journal*, 8, 613–631.

Elfring, G. L. and Schultz, J. R. (1973). Group sequential designs for clinical trials. *Biometrics*, 29(3), 471–477.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Gail, M. H., DeMets, D. L., and Slud, E. V. (1982). Simulation studies on increments of the two sample logrank score test for survival time data, with application to group sequential boundaries. In Johnson, R. and Crowley, J. (eds), *Survival Analysis*, IMS Lecture Notes-Monograph Series, 2. Hayward, California: Institute of Mathematical Statistics

Gange, S. J. and DeMets, D. L. (1996). Sequential monitoring of clinical trials with correlated responses. *Biometrika*, 83, 157–167.

Geary, D. N. (1988). Sequential testing in clinical trials with repeated measurements. *Biometrika*, 75, 311–318.

Gehan, E. A. (1961). The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases*, 13, 346–353.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52, 203–223.

Gu, M. G. and Lai, T. L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *Annals of Statistics*, 19, 1403–1433.

Gu, M. and Ying, Z. (1995). Group sequential methods for survival data using partial score processes with covariate adjustment. *Statistica Sinica*, 5, 793–804.

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics*, 34, 598–611.

Jennison, C. and Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: a review and commentary. *Statistical Science*, 5, 299–317.

Jennison, C. and Turnbull, B. W. (1991). Exact calculations for sequential t, χ^2 and F tests. *Biometrika*, 78, 133–141.

Jennison, C. and Turnbull, B. W. (1997). Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92, 1330–1341.

Kim, K., Boucher, H., and Tsiatis, A. A. (1995). Design and analysis of group sequential logrank tests in maximum duration versus information trials. *Biometrics*, 51, 988–1000.

Kim, K. and DeMets, D. L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*, 74, 149–154.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963–974.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70, 659–663.

Lee, J. W. (1994). Group sequential testing in clinical trials with multivariate observations: a review. *Statistics in Medicine*, 13, 101–111.

Lee, J. W. and DeMets, D. L. (1991). Sequential comparison of changes with repeated measurements data. *Journal of the American Statistical Association*, 86, 757–762.

Lee, J. W. and DeMets, D. L. (1992). Sequential rank tests with repeated measurements in clinical trials. *Journal of the American Statistical Association*, 87, 136–142.

Lee, J. W. and Sather, H. N. (1995). Group sequential methods for comparison of cure rates in clinical trials. *Biometrics*, 51, 756–763.

Lee, S. J., Kim, K., and Tsiatis, A. A. (1996). Repeated significance testing in longitudinal clinical trials. *Biometrika*, 83, 779–789.

Lee, Y. J., Staquet, M., Simon, R., Catane, R., and Muggia, F. (1979). Two-stage plans for patient accrual in phase II cancer clinical trials. *Cancer Treatmetn Reports*, 63, 1721–1726.

Liang, K. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

Liang, K., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorial data (with discussion). *Journal of the Royal Statistical Society*, Series B, 54, 3–40.

Lin, D. (1992). Sequential log rank tests adjusting for covariates with the accelerated life model. *Biometrika*, 79, 523–529.

Lin, D. Y., Wei, L. J., and Ying, Z. (1998). Accelerated failure time models for counting processes. *Biometrika*, 85, 605–618.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50, 163–170.

McCullagh, C. K. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

McPherson, C. K. and Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society*, Series A, 134, 15–26.

Neyman, J. and Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, Series A, 231(694–706), 289–337.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35, 549–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64, 191–199.

Prentice, R. L. and Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics*, 35, 861–867.

Reboussin, D. M., DeMets, D. L., Kim, K., and Lan, K. K. G. (2000). Programs for computing group sequential bounds using the Lan-DeMets method. *Controlled Clinical Trials*, 21, 190–207.

Reboussin, D. M., Lan K. K. G., and DeMets, D. L. (1992). Group sequential testing of longitudinal data. Technical Report 72. Department of Biostatistics, University of Wisconsin-Madison.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Annals of Statistics*, 18, 303–328.

Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77, 485–497.

Scharfstein, D. O., Tsiatis, A. A., and Robins, J. M. (1997). Semiparametric efficiency and its implication on the design and analysis of group sequential studies. *Journal of the American Statistical Association*, 92, 1342–1350.

Schervish, M. J. (1984). Multivariate normal probabilities with error bound (with corrections in 1985). *Applied Statistics*, 33, 81–94.

Schoenfield, L. J., Lachin, J. M., the Steering Committee, and the NCGS Group (1981). Chenodiol for dissolution of gallstones: The National Cooperative Gallstone Study. *Annals of Internal Medicine*, 95, 257–282.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometriak*, 70, 315–326.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10, 1–10.

Slud, E. V. (1984). Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics*, 12(2), 551–571.

Slud, E. V. and Wei, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association*, 77, 862–868.

Sobel, M. and Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics*, 20, 502–522.

Spiessens, B., Lesaffre, E., Verbeke, G., Kim, K., and DeMets, D.L. (2000). An overview of group sequential methods in longitudinal clinical trials. *Statistical Methods in Medical Research*, 9, 497–515.

Spiessens, B., Lesaffre, E., Verbeke, G., and Kim, K. (2002). Group sequential methods for an ordinal logistic random-effects model under misspecification. *Biometrics*, 58, 569–575.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16, 1243–258.

Su, J. Q. and Lachin, J. M. (1992). Group sequential distribution-free methods for the analysis of multivariate observations. *Biometrics*, 48, 1033–1042.

Tarone, R. E. and Ware, J. H. (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, 64, 156–160.

Tsiatis, A. A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*, 68, 311–315.

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association*. 77, 855–861.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics*, 18, 354–372.

Tsiatis, A. A., Boucher, H., and Kim, K. (1995). Sequential methods for parametric survival models. *Biometrika*, 82, 165–173.

Tsiatis, A. A., Rosner, G. L., and Tritchler D. L. (1985). Group sequential tests with censored survival data adjusting for covariates. *Biometrika*, 72, 365–373.

Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, 19, 326–339.

Wei, L. J., Su, J. Q., and Lachin, J. M. (1990a). Interim analyses with repeated measurements in sequential clinical trial. *Biometrika*, 77, 359–364.

Wei, L. J., Ying, Z., and Lin, D. Y. (1990b). Linear regression analysis of censored survival data based on rank tests. *Biometrika*, 77, 845–851.

Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, 45, 939–955.

Wu, M. C. and Lan, K. K. G. (1992). Sequential monitoring for comparison of changes in a response variable in clinical studies. *Biometrics*, 48, 765–779.

SORT 44 (2) July-December 2020, 265-284 DOI: 10.2436/20.8080.02.102

Discrete generalized half-normal distribution and its applications in quantile regression

Diego I. Gallardo[1], Emilio Gómez-Déniz[2] and Héctor W. Gómez[3]

Abstract

A new discrete two-parameter distribution is introduced by discretizing a generalized half-normal distribution. The model is useful for fitting overdispersed as well as underdispersed data. The failure function can be decreasing, bathtub shaped or increasing. A reparameterization of the distribution is introduced for use in a regression model based on the median. The behaviour of the maximum likelihood estimates is studied numerically, showing good performance in finite samples. Three real data set applications reveal that the new model can provide a better explanation than some other competitors.

MSC: 62E10, 62F10, 62P05.

Keywords: Discretizing, generalized half-normal distribution, failure function, health, quantile regression, stochastic order.

1 Introduction

Kemp (2008) introduced a discrete version of the half-normal distribution which, by analogy with the continuous half-normal distribution, is the maximum entropy distribution with specified mean and variance and support on $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Another way of introducing a discrete version of a continuous model is by discretizing it as follows: if $S_Y(x)$ denotes the survival function of a continuous random variable Y with domain in the positive line, the probability mass function (PMF) of its analogue discrete random variable, X, is given by

$$P(X = k) = p_k = S_Y(k) - S_Y(k+1), \quad k \in \mathbb{N}_0. \tag{1}$$

[1] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile, e-mail: diego.gallardo@uda.cl

[2] Department of Quantitative Methods in Economics and TiDES Institute, University of Las Palmas de Gran Canaria, e-mail: emilio.gomez-deniz@ulpgc.es

[3] Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta, Chile, e-mail: hector.gomez@uantof.cl

Received: November 2018
Accepted: June 2020

A classical example is geometric distribution, which can be derived by applying the above discretizing procedure to the negative exponential distribution. Other examples can be found in Nakagawa and Osaki (1975), which obtained the discrete Weibull distribution, Krishna and Singh (2009), the discrete Burr distribution, Gómez-Déniz and Calderín (2011), the discrete Lindley distribution, among many others. This method was also applied by Gómez-Déniz, Vázquez-Polo and García-García (2014) to obtain a discrete version for a generalization of the half-normal distribution based on a skew version of the normal distribution. The resulting discrete distribution differs from that studied in Kemp (2008). The reader can consult the work of Chakraborty (2015) in which different methods and classification are exposed in the discretization procedure of a continuous random variable.

The generalization of the half-normal distribution used in Gómez-Déniz et al. (2014) is based on the idea in Marshal and Olkin (1997). Other generalizations of the half-normal distribution have been proposed in the statistical literature. Here we consider the one in Cooray and Ananda (2008), whose derivation follows from considerations of the relationship between static fatigue crack extension and the failure time of a certain specimen. Its survival function is given by

$$S_Y(x;\sigma,\beta) = 2\Phi\left(-\left(\frac{x}{\sigma}\right)^{\beta}\right), \quad x \geq 0, \tag{2}$$

for some $\sigma,\beta > 0$, where $\Phi(\cdot)$ stands for the cumulative distribution function (CDF) of the standard normal distribution. If a positive random variable Y has survival function (2) we will say that it has a generalized half-normal (GHN) distribution and it will be denoted as $Y \sim GHN(\sigma,\beta)$. The associated discrete version X obtained by applying (1), which will be called the discrete generalized half-normal (DGHN) distribution, has PMF

$$P(X=k;\sigma,\beta) = p(k;\sigma,\beta) = 2\left\{\Phi_\psi\left((k+1)^\beta\right) - \Phi_\psi\left(k^\beta\right)\right\}, \quad x \in \mathbb{N}_0 \tag{3}$$

for some $\sigma,\beta > 0$, where $\psi = \sigma^\beta$ and $\Phi_\sigma(x) = \Phi(x/\sigma)$. If a random variable X taking values on $\mathbb{N}_0$ has PMF (3), we write $X \sim DGHN(\sigma,\beta)$. The new model is different from the one studied in Kemp (2008); for $\beta = 1$ it coincides with that introduced in Gómez-Déniz et al. (2014); for other parameter values, the resulting models are rather different. Figure 1 displays the PMF of X for several parameter. Looking at this figure we see that quite different shapes can be obtained by varying the parameter values.

The discretization of a continuous variable in order to obtain a discrete distribution has been developed with great enthusiasm in recent decades. The simple idea is to start from a continuous random variable that follows a certain probability distribution and for which the distribution function (survival) has a closed form expression. Except for a few occasions (the discretization of the exponential distribution that gives rise to the geometric discrete distribution and the discretization of the Lindley distribution (Gómez-Déniz and Calderín, 2011), the mean and any other superior moment are not obtained in a closed manner.

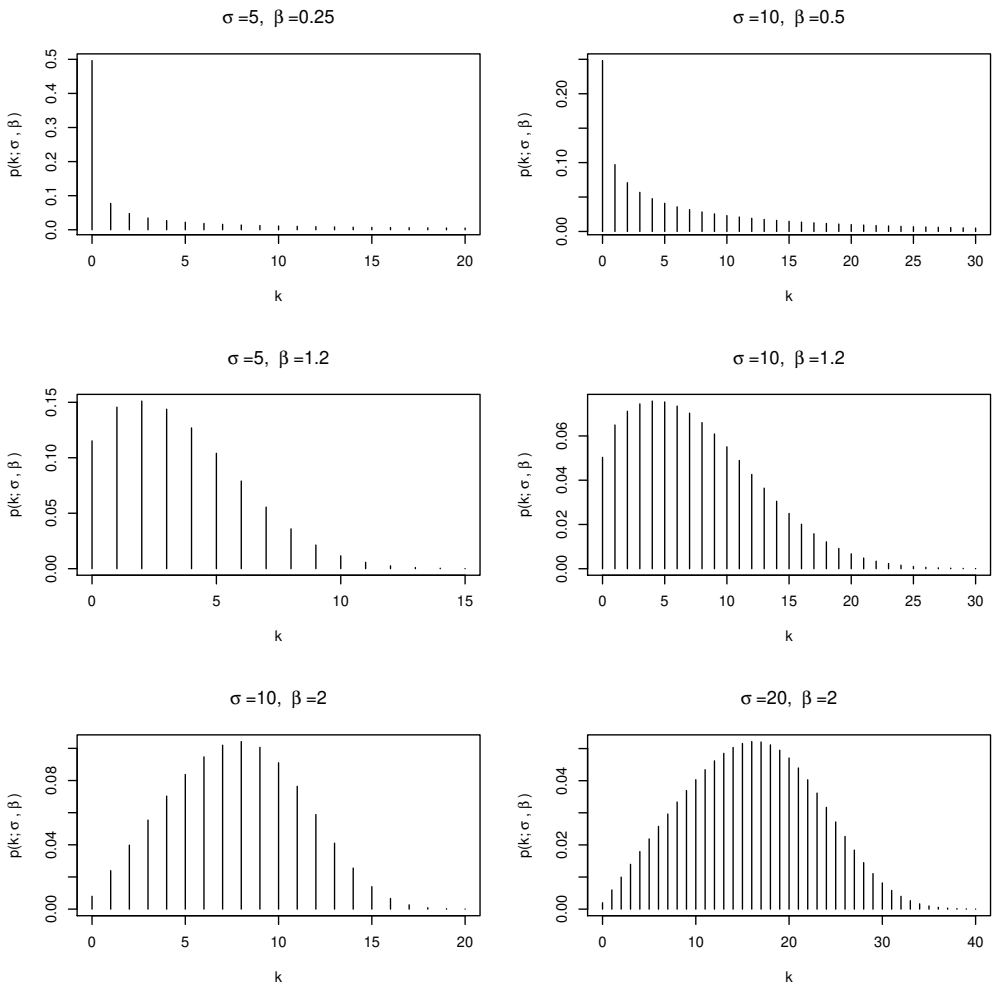


Figure 1: *Some examples of probability mass functions of the DGHN distribution for different values of the parameters β and σ.*

This is a great disadvantage for a researcher who wishes to carry out more in-depth studies on the variable that he wishes to study. For example, a regression study, i.e. explaining the effect that a series of factors can have on the dependent variable, is impossible to perform by ordinary methods.

However, the fact that the distribution function has a closed form makes it easier to calculate the quantile function and therefore to obtain the median. In this case, the initial probability function can be reparametrized as a function of certain parameters, one of which is precisely this quantile, the median. This procedure allows regression analysis to be carried out in a similar way to that traditionally used when trying to explain the mean of the response variable as a function of covariates, which is impossible for the distribution studied here. We therefore propose this line of action in the present work:

we will study the factors that affect the median of the distribution initially verifying that the reparameterization on the median provides a good fit of the data analysed.

The paper is organized as follows. Section 2 gives the expression of some functions associated with the model: the CDF, the survival function and the quantile function; it also explains how to generate random values from the new law, and studies some properties of the model such as unimodality and the fact that its members can be ordered stochastically. Graphical representations show that the family is quite flexible in several senses: it can used to model overdispersed and underdispersed data; it is also seen that the failure function can be decreasing, bathtub shaped or increasing. Section 3 deals with the point estimation of the two parameters. We offer a method of getting a starting point for the optimization problem involved by means of maximum likelihood (ML) estimates. The performance of the ML estimators is studied numerically and shows good behaviour. Finally, Section 4 considers three real data sets. The data are fitted both to the model presented in this paper and to other competitors. The proposed family provides a much better explanation than the other distributions, showing the practical usefulness of the new distribution.

2 Some properties of the discrete generalized half-normal distribution

Let $X \sim DGHN(\sigma,\beta)$, from (3) it readily follows that

$$\frac{p_k}{p_{k-1}} = \frac{\Phi_\psi\left((k+1)^\beta\right) - \Phi_\psi\left(k^\beta\right)}{\Phi_\psi\left(k^\beta\right) - \Phi_\psi\left((k-1)^\beta\right)}, \quad k = 1,2,\ldots,$$

where $p_0 = 1 - 2\Phi_\psi(1)$.

Let $X \sim DGHN(\sigma,\beta)$, from (3) it readily follows that the CDF of X is given by

$$F(k;\sigma,\beta) = 2\Phi_\psi((k+1)^\beta) - 1, \quad k \in \mathbb{N}_0,$$

the survival function of X is

$$S(k;\sigma,\beta) = 2\Phi_\psi(-(k+1)^\beta), \quad k \in \mathbb{N}_0,$$

and the quantile function is given by

$$Q(u;\sigma,\beta) = \left[\sigma\left\{\Phi^{-1}\left(\frac{1+u}{2}\right)\right\}^{1/\beta} - 1\right], \quad u \in (0,1),$$

where $[\cdot]$ denotes the integer part. As a special case, the median is

$$Q(0.5;\sigma,\beta) = \left[\sigma\left\{\Phi^{-1}\left(\frac{3}{4}\right)\right\}^{1/\beta} - 1\right] \approx \left[\sigma\,(0.6745)^{1/\beta} - 1\right]. \tag{4}$$

Because the DGHN distribution is a discrete version of the GHN model, random values can be generated from this distribution as follows:

(i) Generate $u \sim \mathcal{U}(0,1)$.
(ii) Compute $t = \sigma\left(-\Phi^{-1}(u/2)\right)^{1/\beta}$.
(iii) Do $X = [t]$.

2.1 Moments

The moments of X are given by

$$\begin{aligned} E(X^r) &= 2\sum_{k\geq 0} k^r \left\{\Phi_\psi\left((k+1)^\beta\right) - \Phi_\psi\left(k^\beta\right)\right\} \\ &= 2\sum_{k\geq 0}\{(k+1)^r - k^r\}\Phi_\psi\left(-(k+1)^\beta\right). \end{aligned} \tag{5}$$

As $[Y]^r \leq Y^r$, for $r \geq 1$, it follows directly that $E(X^r) < \infty$, $\forall r \in \mathbb{N}$.

In practice, many count data sets exhibit overdispersion and, although less frequently, also underdispersion. Figure 2 shows the value of the quotient $D = Var(X)/E(X)$ when

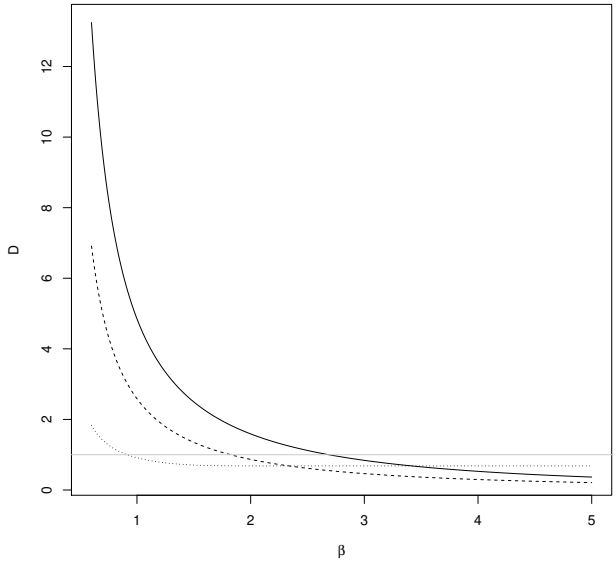


Figure 2: $D = Var(X)/E(X)$ *for* $\sigma = 1$ *(dotted),* $\sigma = 5$ *(dashed) and* $\sigma = 10$ *(solid), the horizontal line* $D = 1$ *is in grey.*

$X \sim DGHN(\sigma,\beta)$ for $\sigma = 1,5,10$ as a function of β. Looking at this figure it can be seen that for each σ the value of D can be greater than, equal to or less than 1 as the value of β increases. In this sense, the new model is quite flexible.

2.2 Mode

Looking at Figure 1 we see that in all cases the PMF is unimodal. Next we show that this is the case for all members in the family. Moreover, we will prove that for $0 < \beta < 1$ the PMF is decreasing. With this aim, we first give a preliminary lemma.

Lemma 1 *If $Y \sim GHN(\sigma,\beta)$ with probability density function $f(x;\sigma,\beta)$, then, as a function of x,*

(a) $f(x;\sigma,\beta)$ is strictly decreasing, if $0 < \beta < 1$, $\forall\sigma > 0$.

(b) $f(x;\sigma,\beta)$ is (strictly) log-concave , if $\beta \geq 1$, $\forall\sigma > 0$.

Proof (a) If $0 < \beta < 1$ then $f(x;\sigma,\beta)$ is proportional to the product of two strictly decreasing functions: $f_1(x) = x^{\beta-1}$ and $f_2(x) = \exp(-0.5x^{2\beta}/\sigma^{2\beta})$; thus it is a strictly decreasing function.

(b) Routine calculations show that $\frac{\partial^2}{\partial x^2} f(x;\sigma,\beta) = -\frac{\beta-1}{x^2} - \frac{\beta(2\beta-1)}{\sigma^{2\beta}} x^{2(\beta-1)}$, which is strictly negative, thus implying the result. ■

Now, we state the following proposition related to the DGHN model.

Proposition 1 *Let $X \sim DGHN(\sigma,\beta)$.*

(a) If $0 < \beta < 1$ and $\sigma > 0$, then $p(k;\sigma,\beta) > p(k+1;\sigma,\beta)$, $\forall k \in \mathbb{N}_0$.

(b) If $\beta \geq 1$ and $\sigma > 0$, then $p(k;\sigma,\beta)^2 \geq p(k-1;\sigma,\beta)p(k+1;\sigma,\beta)$, $\forall k \in \mathbb{N}_0$.

We study separately the two cases: $0 < \beta < 1$ and $\beta \geq 1$.

Proof (a) It is a direct consequence of Lemma 1 (a).

(b) Note that $P(X = k;\sigma,\beta)$ in equation (3) can be written as $P(X = k;\sigma,\beta) = \int_k^{k+1} f(x;\sigma,\beta)dx$. Then, for $\beta \geq 1$, it is a direct consequence of Theorem 2.8. in Dharmadhikari and Joag-Dev (1988) taking $g(x) = f(x;\sigma,\beta)$ (which is log-concave by lemma 1 part b), $\mathcal{B}_n = (0,\infty)$ and $B = (k,k+1) \subseteq \mathcal{B}_n$, that the DGHN distribution is log-concave; the result is immediate. ■

As an immediate consequence of Proposition 1 we state the following.

Corollary 1 *Let $X \sim DGHN(\sigma,\beta)$. X is unimodal. If $0<\beta<1$ the unique mode is attained at $x=0$.*

As commented in Keilson and Gerber (1971), unimodality guarantees that the distribution has all moments, and that the convolution of p_k with any unimodal discrete distribution is also unimodal and log-concave.

2.3 The failure rate function

The failure (or hazard) rate function for the probability function under consideration is given by

$$h(k;\sigma,\beta)=\frac{\Phi_\psi(-k^\beta)}{\Phi_\psi(-(k+1)^\beta)}-1, \quad k\in\mathbb{N}_0.$$

Theorem 9.6 in Dharmadhikari and Joag-Dev (1988) showed that if a random variable is log-concave then it has an increasing failure rate (IFR). Furthermore, Lariviere and Porteus (2001) introduced the concept of generalized failure rate function, defined as $g(k;\sigma,\beta)=k\,h(k;\sigma,\beta)$ for $k\in\mathbb{N}_0$, and showed that the distributions with increasing generalized failure rate (IGFR) have useful applications in operations management (see also Lariviere 2006). It is clear that if a random variable is IFR then it is also IGFR. Accordingly, by the log-concavity of the distribution discussed in Section 2.2, the following result can be established for the discrete generalized half-normal distribution.

Corollary 2 (i) *If $\beta\geq 1$ then the $DGHN(\sigma,\beta)$ distribution is* IFR *and* IGFR .

Figure 3 displays the failure rate function for several parameter values. Looking at this figure, it can be seen that the model is useful for fitting a wide range of shapes: decreasing, bathtub and increasing. Figure 4 shows the different patterns of the failure rate function (IFR, Bathtub and DFR) accordingly to the values of σ and β. We highlight that for $0<\beta\leq 1/2$ the model seems to be DFR, whereas for $1/2<\beta<1$ the behaviour of the failure rate also depends on σ.

The next proposition shows the limit of the failure rate for $k\to+\infty$.

Proposition 2 *Let $X\sim DGHN(\sigma,\beta)$. Therefore, the failure rate satisfies*

$$\lim_{k\to\infty} h(k;\sigma,\beta)=\begin{cases} 0 & \text{if } 0<\beta<1/2,\\ \exp\left(\frac{1}{2\sigma}\right)-1 & \text{if } \beta=1/2,\\ \infty & \text{if } \beta>1/2.\end{cases}$$

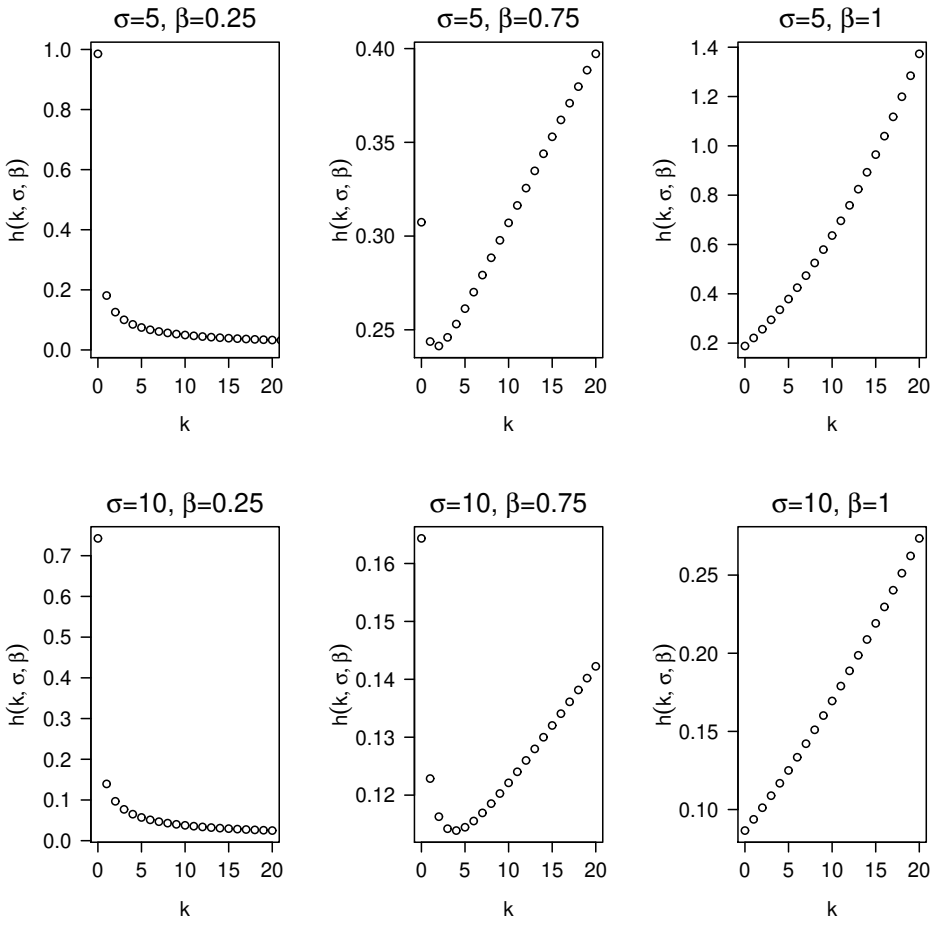


Figure 3: *Failure rate function for several parameter values.*

Proof Using the L'Hôpital rule and the continuity of the limit, we have

$$\lim_{k\to\infty} h(k;\sigma,\beta) = \lim_{k\to\infty}\left(\frac{k}{1+k}\right)^{\beta-1} \exp\left\{-\frac{1}{2\sigma^{2\beta}}\lim_{k\to\infty}\frac{\left[1-\left(1+\frac{1}{k}\right)^{2\beta}\right]}{k^{-2\beta}}\right\} - 1$$

Applying the L'Hôpital rule again in the second limit, we have

$$\begin{aligned}\lim_{k\to\infty} h(k;\sigma,\beta) &= \exp\left\{\frac{1}{2\sigma^{2\beta}}\lim_{k\to\infty}\frac{\left(1+\frac{1}{k}\right)^{2\beta-1}}{k^{-2\beta+1}}\right\} - 1 \\ &= \exp\left\{\frac{1}{2\sigma^{2\beta}}\lim_{k\to\infty}(1+k)^{2\beta-1}\right\} - 1.\end{aligned}$$

The result is obtained separating the cases $0 < \beta < 1/2$, $\beta = 1/2$ and $\beta > 1/2$. ■

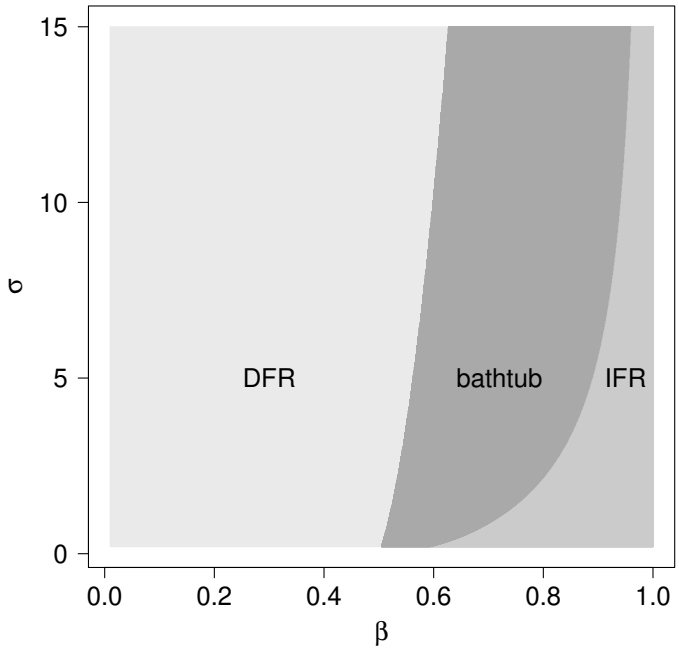


Figure 4: *Shapes for the failure rate of DGHN*(σ,β) *for* $0<\beta<1$.

2.4 Stochastic orderings

This subsection shows that the members of the new model can be stochastically ordered according to the parameter values. With this aim, we first recall the following definition:

Definition 1 *Let X_1 and X_2 be two random variables with distribution functions F_1 and F_2, respectively. Then X_1 is said to be stochastically smaller than X_2, denoted by $X_1 \leq_{st} X_2$, if $F_1(x) \geq F_2(x)$ for all x.*

The DGHN family can be ordered in the following way.

Proposition 3 *(a) Let $X_1 \sim DGHN(\sigma,\beta_1)$ and $X_2 \sim DGHN(\sigma,\beta_2)$, for some σ,β_1, $\beta_1>0$. Then, $X_2 \leq_{st} X_1$ if and only if $\beta_1 \geq \beta_2$.*

(b) Let $X_1 \sim DGHN(\sigma_1,\beta)$ and $X_2 \sim DGHN(\sigma_2,\beta)$, for some $\sigma_1,\sigma_2,\beta>0$. Then, $X_2 \leq_{st} X_1$ if and only if $\sigma_1 \geq \sigma_2$.

Proof (a) Let $\psi_i=\sigma^{\beta_i}$, $i=1,2$. We have $X_2 \leq_{st} X_1$ if and only if $P(X_2 \geq x) \leq P(X_1 \geq x)$ for all $x \in \mathbb{N}_0$ if and only if $2\Phi_{\psi_2}(-(x+1)^{\beta_2}) \leq 2\Phi_{\psi_1}(-(x+1)^{\beta_1})$ for all $x \in \mathbb{N}_0$ if and only if $\beta_1 \geq \beta_2$.

(b) The result can be shown using a similar argument to (a). ■

The following corollary is a consequence of Proposition 3.

Corollary 3 *(i) If* $X_1 \sim DGHN(\sigma,\beta_1)$ *and* $X_2 \sim DGHN(\sigma,\beta_2)$*, with* $\beta_1 \geq \beta_2$*, then* $E(X_2^r) \leq E(X_1^r)$*, for all* $r > 0$*.*

(ii) If $X_1 \sim DGHN(\sigma_1,\beta)$ *and* $X_2 \sim DGHN(\sigma_2,\beta)$*, with* $\sigma_1 \geq \sigma_2$*, then* $E(X_2^r) \leq E(X_1^r)$*, for all* $r > 0$*.*

3 Point estimation

3.1 Without covariates

Let $X_1,\ldots,X_n$ be independent and identically distributed (IID) from $X \sim DGHN(\sigma,\beta)$, and let the observed values be denoted by $x_1,\ldots,x_n$. The log-likelihood function for (σ,β) is

$$\ell(\sigma,\beta) = n\log(2) + \sum_{i=1}^{n} \log\left\{\Phi_\psi\left((x_i+1)^\beta\right) - \Phi_\psi\left(x_i^\beta\right)\right\}. \tag{6}$$

The derivatives of the log-likelihood function are

$$\frac{\partial}{\partial\sigma}\ell(\sigma,\beta) = -\frac{\beta}{\sigma}\sum_{i=1}^{n} \frac{\phi\left(\frac{(x_i+1)^\beta}{\sigma^\beta}\right)\frac{(x_i+1)^\beta}{\sigma^\beta} - \phi\left(\frac{x_i^\beta}{\sigma^\beta}\right)\frac{x_i^\beta}{\sigma^\beta}}{\Phi_\psi\left((x_i+1)^\beta\right) - \Phi_\psi\left(x_i^\beta\right)}, \tag{7}$$

$$\frac{\partial}{\partial\beta}\ell(\sigma,\beta) = \frac{1}{\sigma^\beta}\sum_{i=1}^{n} \frac{\phi\left(\frac{(x_i+1)^\beta}{\sigma^\beta}\right)(x_i+1)^\beta\log\left(\frac{x_i+1}{\sigma}\right) - \phi\left(\frac{x_i^\beta}{\sigma^\beta}\right)x_i^\beta\log\left(\frac{x_i}{\sigma}\right)}{\Phi_\psi\left((x_i+1)^\beta\right) - \Phi_\psi\left(x_i^\beta\right)}. \tag{8}$$

The ML estimates of the parameters satisfy the system that results from equating to 0 in equations (7) and (8). Nevertheless, since this system does not have an explicit solution, in order to obtain the ML estimates it is preferable to maximize function (6). This can be carried out, for example, by using the BFGS algorithm available in the `optim` function of the R programming language (R Core Team, 2016). The BFGS algorithm requires a starting point, which must be inside the feasible region. The estimators obtained from equating any two observed frequencies to their theoretical values can be used as the starting point. For example, if $\widehat{p}_i$ denotes the observed frequency of the value i, for $i = 0,1$ (the zero-frequency and the one-frequency method), the system is

$$\widehat{p}_0 = 2\Phi_\psi(1) - 1 \qquad \text{and} \qquad \widehat{p}_1 = 2\left\{\Phi_\psi\left(2^\beta\right) - \Phi_\psi(1)\right\}.$$

The solutions for ψ and β obtained from the above equations are

$$\tilde{\psi} = \left[\Phi^{-1}\left(\frac{1+\widehat{p}_0}{2}\right)\right]^{-1} \qquad \text{and} \qquad \tilde{\beta} = \frac{\log\tilde{\psi} + \log\Phi^{-1}\left(\widehat{p}_1/2 + \Phi\left(1/\tilde{\psi}\right)\right)}{\log 2}.$$

Therefore, the solution for σ is $\tilde{\sigma} = \tilde{\psi}^{1/\tilde{\beta}}$.

In order to assess numerically the performance of the ML estimates, a simulation study was carried out. Below we describe the study and summarize the results obtained. For several values of the parameters ($\beta = 0.8$, 1.0, 1.3 and $\sigma = 1$, 5) and sample sizes ($n = 30$, 50, 100) 1000 random samples were generated. In each case, the ML estimates of β and σ were computed, as well as their standard error based on the hessian matrix of the model. Table 1 reports the bias, the root of the mean squared error ($\sqrt{MSE}$) and the coverage probability (CP) of the 95% level interval obtained from the asymptotic normality of the ML estimates. As expected, the bias and the $\sqrt{MSE}$ decrease as the sample size increases. Also as expected, the closeness of the CP to its nominal value increases as the sample size increases.In all cases the empirical coverages is quite close to 0.95.

Table 1: *Results for the ML estimates in the DGHN model.*

			$n=30$			$n=50$			$n=100$		
β	σ		bias	$\sqrt{MSE}$	CP	bias	$\sqrt{MSE}$	CP	bias	$\sqrt{MSE}$	CP
0.8	1	$\widehat{\beta}$	0.157	0.443	0.970	0.067	0.249	0.962	0.030	0.125	0.954
		$\widehat{\sigma}$	0.006	0.211	0.955	0.001	0.166	0.952	0.003	0.117	0.950
	5	$\widehat{\beta}$	0.040	0.150	0.955	0.027	0.113	0.950	0.013	0.075	0.952
		$\widehat{\sigma}$	-0.007	0.901	0.926	-0.007	0.695	0.933	-0.003	0.490	0.940
1	1	$\widehat{\beta}$	0.304	0.679	0.970	0.156	0.468	0.971	0.047	0.210	0.961
		$\widehat{\sigma}$	-0.002	0.166	0.969	0.001	0.131	0.963	-0.002	0.094	0.953
	5	$\widehat{\beta}$	0.055	0.190	0.949	0.030	0.137	0.952	0.015	0.092	0.954
		$\widehat{\sigma}$	-0.017	0.708	0.931	-0.005	0.550	0.937	-0.009	0.387	0.944
1.3	1	$\widehat{\beta}$	0.648	0.948	0.975	0.520	0.868	0.975	0.266	0.620	0.975
		$\widehat{\sigma}$	-0.002	0.118	0.980	-0.001	0.094	0.980	0.001	0.070	0.957
	5	$\widehat{\beta}$	0.071	0.237	0.958	0.039	0.175	0.948	0.021	0.116	0.957
		$\widehat{\sigma}$	-0.023	0.549	0.926	-0.020	0.427	0.932	-0.006	0.299	0.947

3.2 Estimation in a DGHN regression model

Unfortunately, the mean of the DGHN has a complicated form (see equation (5)). For this reason, an alternative way to use this model in a regression context is through the median (see equation (4)). Let $Q_{0.5}$ be the median of the model. The pmf of the model

with reparametrization based on $Q_{0.5}$ and β is given by

$$p_k = 2\sum_{j=0}^{1}(-1)^j \Phi\left(\tau\left(\frac{k+j}{1+Q_{0.5}}\right)^{\beta}\right), \quad k = 0,1,2,\ldots \tag{9}$$

where $\tau = 0.674489$.

A common specification for $Q_{0.5}$ is exponential, ensuring the non–negativity of this parameter. That is,

$$\log Q_{0.5i} = \sum_{s=1}^{\kappa} x_{is}\gamma_s, \quad i = 1,\ldots,t,$$

where $x_{i1}, x_{i2}, \ldots, x_{i\kappa}$ are covariates and $\gamma_1, \gamma_2, \ldots, \gamma_\kappa$ are unknown regression coefficients. The log-likelihood for the vector (γ, β) is

$$\ell(\gamma,\beta) = n\log 2 + \sum_{i=1}^{n}\log\left\{\Phi\left(\tau\left(\frac{k}{1+Q_{0.5i}}\right)^{\beta}\right) - \Phi\left(\tau\left(\frac{k+1}{1+Q_{0.5i}}\right)^{\beta}\right)\right\}. \tag{10}$$

Again, the mle of (γ, β) can be obtained maximizing (10) in relation to them.

4 Applications

This section presents applications to three real data sets.

4.1 An application in ecology

This data set (Kulasekera and Tonkyn, 1992 and Table **2** here) consists of the number of weevil eggs laid per bean and contains 193 observations.

Table 2: *Number of weevil eggs laid per bean*

Number / bean	0	1	2	3	Total
Obs. Freq.	5	68	88	32	193

To analyse the data we considered the model proposed in this paper, comparing it to the models in Kemp (2008), Gómez-Déniz et al. (2014) and in Kulasekera and Tonkyn (1992) (denoted as Kula in the tables). ML estimators of the parameters for each model are shown in Table 3. This table also shows the value of the maximized log-likelihood, L, and the Akaike information criterion, Akaike (1974), defined as $\mathrm{AIC} = 2r - 2\log L$, where r is the number of parameters. As is well-known, the model with lower AIC is preferred. Therefore, according to this criterion, the proposed model provides a better fit than the other laws. To illustrate the performance of the DGHN model for this data,

we estimate the probability of the events $X = 0$, $X = 1$, $X = 2$, $X = 3$ and $X \geq 4$ for all the models with their respective 95% confidence intervals based on the delta method (we exclude the estimations provided by Kulasekera and Tonkyn (1992) because their intervals are very wide). Results are presented in Table 4. Note that the DGHN model is the only one for which the confidence intervals always include the observed frequencies. Therefore, the proposed distribution may be an attractive alternative to models for data taking values in $\mathbb{N}_0$.

Table 3: *Model ML estimates and standard errors (in parentheses).*

	Kemp	Gómez-Déniz et al. (2014)	DGHN	Kula
	$\widehat{\theta} = 12.9970\ (15.2697)$	$\widehat{\alpha} = 54.1196\ (0.2091)$	$\widehat{\beta} = 2.8873\ (3.0927)$	$\widehat{\alpha} = 11.0943\ (13.8496)$
	$\widehat{q} = 0.1393\ (0.0490)$	$\widehat{\sigma} = 1.0860\ (0.0562)$	$\widehat{\sigma} = 2.6519\ (0.0251)$	$\widehat{q} = 0.0125\ (0.0057)$
L	–223.956	–222.9054	–218.7891	–221.9045
AIC	451.9119	449.8108	441.5782	447.8090

Table 4: *Estimated probabilities for $P(X = k), k = 0,1,2,3$, and $P(X \geq 4)$ and their 95% confidence intervals (CI).*

	$X=0$		$X=1$		$X=2$		$X=3$		$X\geq 4$	
model	point	95% CI	point	95% CI	point	95% CI	point	95% CI	point	95% CI
Kemp	0.022	(0.011,0.034)	0.291	(0.249,0.333)	0.527	(0.476,0.577)	0.133	(0.093,0.173)	0.005	(0.000,0.009)
Gómez-Déniz et al. (2014)	0.061	(0.039,0.084)	0.280	(0.235,0.324)	0.522	(0.463,0.581)	0.131	(0.090,0.172)	0.006	(0.001,0.011)
DGHN	0.048	(0.025,0.069)	0.294	(0.253,0.356)	0.505	(0.449,0.561)	0.152	(0.106,0.199)	0.001	(0.000,0.003)
observed	0.026		0.352		0.456		0.166		0.000	

4.2 A real application in the health framework

Since the seminal work of Koenker and Bassett (1978) quantile regression has attracted much research, particularly in recent years, probably due to the help of computers. This technique allows a natural generalization of the generalized linear models for certain well-known robust estimators of location. The methodology we propose in this Section is simple and, enables us to explain the median by the effects of covariate factors, as discussed in Section 3.2.

Many authors in the literature have focused on the factors that affect the mean of the dependent variable under study. The proposal presented here is based on studying the factors that can affect the median of the dependent variable. As far as we know, there are few studies in the theoretical or applied statistical literature of regression of quantiles for a discrete variable (parametric model).

A common specification for the median parameter, $Q_{0.5}$, is exponential, ensuring the non–negativity of the parameter. That is,

$$\log Q_{0.5} = \sum_{s=1}^{\kappa} x_{is}\gamma_s, \quad i = 1,\ldots,t,$$

obtaining the conventional log-linear model such that $Q_{0.5} = \exp\left\{\gamma^\top x\right\}$, where x is the vector of covariates and γ is an unknown vector of regression coefficients.

The marginal effect, which reflects the variation of the conditional median due to a one-unit change in the jth covariate $(j = 1,\ldots,\kappa)$, has a similar consideration to that in generalized linear models. For indicator variables such as $x_{\kappa i}$ which takes only the values 0 or 1, the marginal effect is $\delta_j = Q_{0.5}(k_i|x_j = 1, x_1,\ldots,x_\kappa)/Q_{0.5}(k_i|x_j = 0, x_1,\ldots,x_\kappa) \approx \exp(\beta_j)$, $i = 1,\ldots,n$; $j = 1,\ldots,\kappa$. Therefore, the conditional median is $\exp(\beta_j)$ times larger if the indicator variable is one rather than zero.

For the present purpose we used data obtained from the 1977-78 Australian Health Survey, a well-known data set previously studied by Cameron and Trivedi (1998); see also Cameron and Trivedi (1986). This data set can be downloaded from the web page

`http://cameron.econ.ucdavis.edu/racd/racddata.html`

Details of this data source can also be consulted in the "Ecdat" R (data(DoctorAUS)) package. The data set consists of 5190 elements with fifteen variables. The variable ILLNESS, the number of illnesses in past 2 weeks is taken as the dependent variable. The minimum value of this variable is 0, the maximum value 5 and the median is 1. A different count variable could be taken as the dependent variable if another study were required. Fundamentally, the convenience of this approach is based on the fact that by testing all the count variables appearing in the data, the variable ILLNESS presents a median different from zero and a larger index of dispersion.

In our study, CHCOND (chronic condition) is not considered, and INSURANCE (medlevy : medibanl levy, levyplus: private health insurance, freepoor: government insurance due to low income, freerepa : government insurance due to old age disability or veteran status) is converted into three dichotomous variables, FREEPOR, FREEREPA AND LEVYPLUS. Therefore, MEDLEVY is the reference variable.

Descriptive statistics on the variables in this dataset are given in Cameron and Trivedi (1986, p.68) (see Table 3.2 in this work). In our study the following distributions were also considered for comparison purposes: a Poisson (P) distribution with parameter $\beta > 0$; a negative binomial (NB) distribution with parameters $\beta > 0$ and mean $q > 0$; a generalised Poisson (GP) distribution with parameters $\beta > 0$ and mean $q > 0$ and of course the proposed distribution studied here. Among the various parameterisations of the generalized Poisson distribution, we used the one described in Consul and Famoye (1992).

Tables 5 and 6 show the estimation in the case of non including and including covariates, respectively. Again, in view of the maximum value of the logarithm of the likelihood function, the proposed distribution studied here is superior to the remainders. We estimated the two parameters, β and $q = Q_{0.5}$ by maximizing directly the log-likelihood function given by $L = \sum_{i=1}^{n} \log p_{k_i}$. We also show the value obtained for the Akaike Information Criterion (AIC). (Note that AIC $= 2(k - L)$, where k is the number of model

parameters and L is the maximum value of the log–likelihood function). The goodness of fit is also corroborated by looking at the graph shown in Figure 5, in which it can be observed that the model seems to be a reasonable choice for the given data.

Table 5: *Coefficient estimates and p-values for the different models considered without covariates.*

	P		NB		GP		DGHN	
Parameter	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
$\widehat{\beta}$	1.431	0.000	3.801	0.000	0.120	0.000	0.082	0.000
$\widehat{q}$			1.431	0.000	1.432	0.000	0.015	0.000
L	-8390.942		-8264.408		-8266.708		-8255.156	
AIC	16783.90		16532.80		16537.40		16514.30	

Table 6: *Coefficient estimates and p-values for the different models considered with covariates. The cases of P, NB and GP correspond to maximizing the mean link and the GHN to maximizing the median*

	P		NB		GP		DGHN	
Variable	Estimate	$\Pr > \|t\|$	Variable	$\Pr > \|t\|$	Variable	$\Pr > \|t\|$	Variable	$\Pr > \|t\|$
SEX	0.022	0.259	0.021	0.419	0.021	0.407	0.013	0.750
AGE	0.151	0.026	0.143	0.081	0.142	0.080	0.367	0.003
INCOME	-0.125	0.000	-0.125	0.001	-0.125	0.001	-0.186	0.002
HSCORE	0.082	0.000	0.084	0.000	0.084	0.000	0.126	0.000
DOCTORCO	0.043	0.000	0.045	0.000	0.045	0.000	0.060	0.002
NONDOCCO	0.009	0.253	0.008	0.415	0.008	0.384	0.000	0.962
HOSPADMI	-0.014	0.433	-0.012	0.614	-0.012	0.611	-0.011	0.716
HOSPDAYS	0.000	0.475	0.000	0.655	0.000	0.638	0.001	0.463
MEDECINE	0.071	0.000	0.072	0.050	0.072	0.000	0.095	0.000
PRESCRIB	0.077	0.000	0.078	0.037	0.078	0.000	0.097	0.000
NONPRESC	0.103	0.000	0.105	0.007	0.105	0.000	0.154	0.000
FREEPOR	0.008	0.610	0.009	0.936	0.009	0.720	-0.040	0.209
FREEREPA	0.103	0.003	0.107	0.015	0.107	0.011	0.136	0.044
LEVYPLUS	0.008	0.610	0.009	0.936	0.009	0.720	0.049	0.128
CONSTANT	-0.064	0.084	-0.068	0.122	-0.069	0.114	-0.968	0.000
$\widehat{\beta}$			38.373	0.053	0.013	0.053	1.213	0.000
L	-7590.674		-7588.737		-7588.696		-7759.528	

As can be seen, most of the covariates considered are statistically significant except SEX, NONDOCCO, HOSPADMI, HOSPDAYS, FREEPOR and LEVYPLUS in all the models used. Observe that the sign of the regressors coincides for all the models.

It can be seen that the maximum value of the log-likelihood function is lower in the case of the quantile regression although the estimates are similar in terms of sign and significance. This is not surprising since the link used affects the mean in classical models and the median in the distribution studied here. Thus from our point of view, the model is viable for cases in which classical distributions provide a poor fit of the variable to be studied, as will be seen in the last example provided in the next subsection.

The different models considered were analysed using the BFGS algorithm (Broyden, Fletcher, Goldfarb and Shanno), with RATS and Mathematica (Wolfram) software, for

both the inflated and the non-inflated models. In all of the models considered, the convergence of the algorithm is extremely fast. In general, the algorithm converged in fewer than 30 iterations.

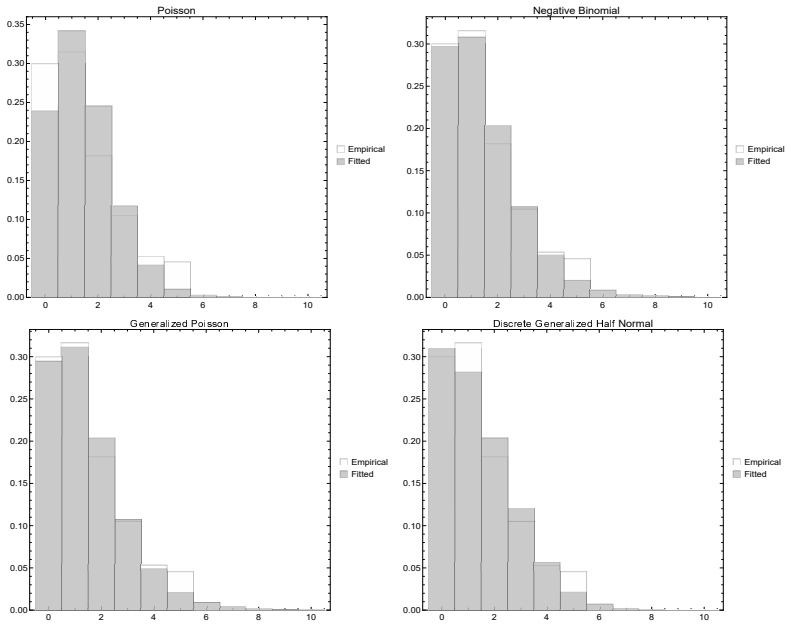


Figure 5: *Empirical and fitted data for the number of illness in the past two weeks.*

4.3 An actuarial application

Usually in automobile insurance rate-making the target is to estimate the probability of a claim in order to compute a premium according to a premium calculation principle. In this example we consider a dataset of Swedish third party automobile insurance claims which is well-known in the actuarial literature. Some of the most important factors of claim frequency will be taken into account. The variable kilometres (Km) is the kilometers travelled by a vehicle, here grouped into seven categories (category 1, less than 1000 km per year, category 2, 1000-15000 km per year, etc.); Zone gives the graphic zone, also grouped into seven categories; Bonus is a variable representing the driver claim record grouped into seven categories; Insured starts in the class 1 and is moved up one class, to a maximum of 7, for each year in which there is no claim; finally, Make represents the type of vehicle (nine specified makes of car). The dependent variable is the Number of claims. More details can be seen in Frees (2010).

For comparison purposes we have considered the Poisson and the negative binomial distributions, which are very widely used in the actuarial context, to fit the number of

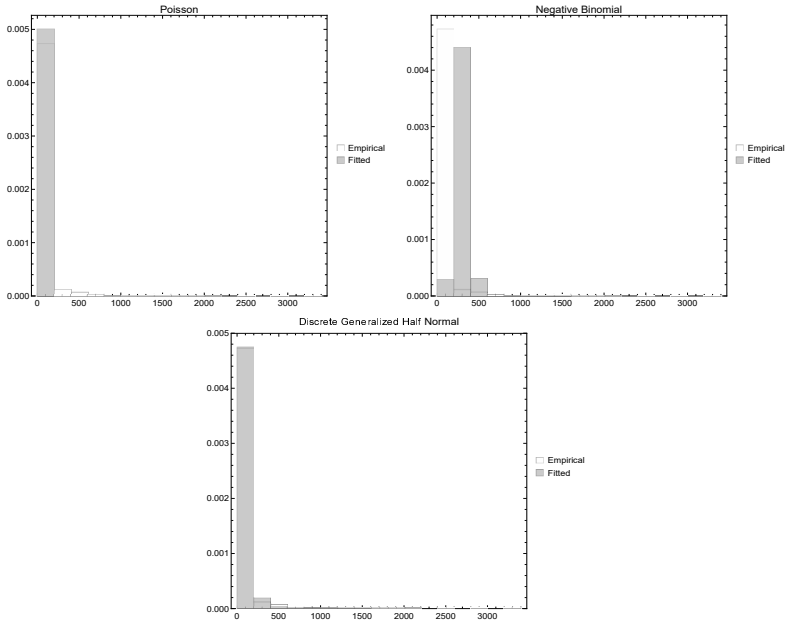


Figure 6: *Empirical and fitted data for Swedish automobile claims.*

claims. The values of the maximum of the log-likelihood function for these models are -221571.00 and -93806.541, respectively, compared to -8920.57 for the distribution proposed here. Therefore, the proposed distribution is very much superior to the others. The estimated values of the parameters are $\widehat{\beta} = 0.324\,(0.005)$ and $\widehat{q} = 1.571\,(0.197)$ (standard errors in parentheses).

Figure 6 shows the empirical and fitted distributions obtained using Poisson, negative binomial and our proposed distribution. This graphic confirms the superiority of the proposed distribution over the others.

Using a similar idea to that proposed in Heras, Moreno and Vilar-Zanón (2018), we have used the covariates explained above in order to explain the median of the dependent variable given by the number of claims. The results are shown in Table 7. As we can see the value of the maximum of the log-likelihood function has been much reduced.

It can be seen that all the variables are highly significant, and the signs (see Frees , 2010) are similar to those of the classical regression model when the Number of claims is considered as the dependent variable, except for the covariate Km; when this last is studied in detail, the interpretation is observed to be similar; the covariable Km takes values from 1 to 5, increasing with the number of kilometers traveled by the insured. The negative value of the regressor indicates that the greater the number of kilometers traveled, the smaller will be the value of the median. The insured will have better insurance terms than justified by his claim record (because he has travelled more kilometers).

Finally, the premium for this automobile insurance portfolio, which is not computed here, can be obtained by using the quantile principle used by Heras et al. (2018).

Table 7: *Parameter estimates from the new count distribution using quantile (median) regression*

Parameter	Estimate	S.E.	t–Wald	$\Pr > \|t\|$
km	–0.536	0.035	15.096	0.000
zone	–0.394	0.026	14.802	0.000
bonus	0.253	0.020	12.382	0.000
make	0.289	0.015	19.251	0.000
$\widehat{\beta}$	0.410	0.006	58.834	0.000
constant	2.485	0.183	13.581	0.000
$L = -8516.35$				
AIC = 17044.70				

Conclusions

This work introduces the discrete version of the continuous GHN distribution. We have presented its most important probabilistic properties. Parameter estimation was approached by maximum likelihood. Using three applications to real data sets, we have shown that the discrete generalized half-normal distribution proposed in this work provided a better fit than other extensions of the discrete half-normal model, illustrating that the model is competitive with other discrete models depending on two parameters.

One of the disadvantages of the discretization of a continuous variable is that the average does not appear expressed in a closed form allowing simple reparameterization of the distribution in order to incorporate covariables. However, as noted, this drawback can be avoided by carrying out quantile regression (the median in our case). This is possible due to the fact that the discretization is carried out from the distribution function, which has a simple, closed expression. This particularity has been incorporated into this work with an application in the health scenario, which take into account the fact that on many occasions the median is a more intuitive, manageable and practical characteristic than the mean.

Acknowledgments

The authors are grateful to two anonymous Referees as well as the Associate Editor for their contributions, which have improved the presentation of this work.

The research of D. I. Gallardo was supported by FONDECYT 11160670 (Chile). The research of H. W. Gómez was supported by MINEDUC-UA project, code ANT 1755 (Chile). EGD's work was partially funded by grant ECO2017–85577–P (Ministerio de Economía, Industria y Competitividad. Agencia Estatal de Investigación)). EGD also acknowledges the Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta (Chile) for their special support, as part of this work was done while EGD was visiting this University in 2018.

References

Akaike, H. (1974). A new look at statistical model identification, *IEEE Transactions on Automatic Control*, 19, 716-723.

Cameron, A.C. and Trivedi, P.K. (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics*, 1, 29-54.

Cameron, C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Chakraborty, S. (2015). Generating discrete analogues of continuous probability distributions-A survey of methods and constructions. *Journal of Statistical Distributions and Applications*, 2, 6.

Consul, P.C. and Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, 21, 89-109.

Cooray, K. and Ananda, M. (2008). A Generalization of the Half-Normal Distribution with Applications to Lifetime Data. *Communications in Statistics - Theory and Methods*, 37, 1323-1337.

Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, convexity and applications. Probability and mathematical statistics*. Academic Press Inc, Boston

Frees, E.W. (2010). *Regression Models with Actuarial and Financial Applications*. Cambridge University Press

Gómez-Déniz, E. and Calderín, E. (2011). The discrete Lindley distribution: properties and applications. *Journal of Statistical Computation and Simulation*, 81, 1405-1416.

Gómez-Déniz, E., Vázquez-Polo, F.J. and García-García, V. (2014). A discrete version of the half-normal distribution and its generalization with applications. *Statistical Papers*, 55, 497-511.

Heras, A., Moreno, I. and Vilar-Zanón, J.L. (2018). An application of two-stage quantile regression to insurance ratemaking. *Scandinavian Actuarial Journal*, 9, 753-769.

Keilson, J. and Gerber, H. (1971). Some results for discrete unimodality. *Journal of the American Statistical Association*, 66, 386-389.

Kemp, A.W. (2008). The discrete half-normal distribution. In: Birkhä (Ed) *Advances in mathematical and statistical modeling*, pp. 353-365.

Koender, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33-50.

Krishna, H. and Singh, P. (2009). Discrete Burr and discrete Pareto distributions. *Statistical Methodology*, 6, 177-188.

Kulasekera, K.B. and Tonkyn, D.W. (1992). A new discrete distribution with applications to survival, dispersal and dispersion. *Communications in Statistics - Simulation and Computation*, 21, 499-518.

Lariviere, M.A. and Porteus, E.L. (2001). Selling to the newsvendor: An analysis of price-only contracts. *M&SOM*, 3, 293-305.

Lariviere, M.A. (2006). A note on probability distributions with increasing generalized failure rates. *Operations Research*, 54 602-604.

Marshall, A.W. and 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.

Nakagawa, T. and Osaki, S. (1975). The discrete Weibull distribution. *IEEE Transactions on Reliability*, 24, 300-301.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

SORT 44 (2) July-December 2020, 285-310 DOI: 10.2436/20.8080.02.103

A simheuristic algorithm for time-dependent waste collection management with stochastic travel times

Aljoscha Gruler[1], Antoni Pérez-Navarro[1,*], Laura Calvet[1,2] and Angel A. Juan[1]

Abstract

A major operational task in city logistics is related to waste collection. Due to large problem sizes and numerous constraints, the optimization of real-life waste collection problems on a daily basis requires the use of metaheuristic solving frameworks to generate near-optimal collection routes in low computation times. This paper presents a simheuristic algorithm for the time-dependent waste collection problem with stochastic travel times. By combining Monte Carlo simulation with a biased randomized iterated local search metaheuristic, time-varying and stochastic travel speeds between different network nodes are accounted for. The algorithm is tested using real instances in a medium-sized city in Spain.

MSC: 90B06, 68U20, 68T20, 90C59.

Keywords: Waste collection management, vehicle routing problem, stochastic optimization, simheuristics, biased randomization, case study.

1 Introduction

Due to its high operational costs and numerous related negative externalities such as air pollution, noise, and traffic congestion, waste management is among the most important public services (Strand, Syberfeldt and Geertsen, 2020). The complete process of collecting and disposing different types of garbage is a complex task shaped by various optimization problems related to facility location, clustering of service territories, and vehicle routing (Ghiani et al., 2014). Considering rising population numbers in urban areas around the world, especially waste collection processes need to be organized in an efficient manner in order to ensure a sustainable, cost-efficient, and

* *Corresponding author:* Antoni Pérez-Navarro, e-mail: aperezn@uoc.edu

[1] IN3 – Computer Science, Multimedia and Telecommunication Department, Universitat Oberta de Catalunya, 08018 Barcelona, Spain.

[2] Business Area, International University of Valencia, 46002 Valencia, Spain.

Received: July 2019
Accepted: June 2020

citizen-friendly metropolitan garbage collection (Bing et al., 2016). The waste collection problem (WCP) is a rich extension of the well-known vehicle routing problem (VRP) with the aim of minimizing a certain objective function, e.g.: distances, travel times, CO_2 emissions, etc. (Kim, Kim and Sahoo, 2006). Problem inputs include a set of waste containers that hold a positive amount of waste, which has to be collected from a number of capacitated garbage collection vehicles located at a central depot. Moreover, the problem setting includes one or more landfills at which collected waste is disposed if a vehicle is full or before it returns to the central depot.

Given the practical nature of the WCP, realistic problem instances discussed in the literature typically include several hundred waste containers and several constraints related to maximum route travel times, driver lunch breaks, time windows, etc. (Benjamin and Beasley, 2010; Buhrkal, Larsen and Ropke, 2012). This imposes certain limits on the use of exact methods to solve this NP-hard problem, calling for the application of metaheuristic algorithms that are able to generate near-optimal solutions to large-scaled and realistic WCP settings in calculation times of only a few seconds or minutes. However, most metaheuristic solving methodologies still make simplifying assumptions about the nature of input variables. On the one hand, most routing optimization frameworks assume travel times between different network nodes to be static over time. Especially in the context of daily collection of waste, this is an unrealistic assumption due to the natural time dependency of edge traversing duration and vehicle velocities (Gendreau, Ghiani and Guerriero, 2015). On the other hand, a frequent drawback of many solving approaches is that they do not consider uncertainty in input variables. In the context of vehicle routing, information regarding travel times, demands, or customers themselves is typically not perfectly known in advance. Indeed, they are more likely to be of stochastic or even dynamic nature (Pillac et al., 2013; Ritzinger, Puchinger and Hartl, 2016).

Figure 1 illustrates the effects of time-dependent and stochastic travel speeds. Given the distance of traversing any edge in a routing problem, the travel duration to pass this edge can be calculated as the quotient of travel distance and the expected vehicle speed. In time-dependent routing scenarios, driving velocities vary according to different time periods within the route planning horizon. Apart from the expected travel speeds, realistic problem settings should also consider travel time variances due to different levels of planning uncertainty. The effects of different travel time assumptions are highlighted as optimistic and pessimistic vehicle speeds below, showing that variances in vehicle velocities can significantly impact the necessary time to visit a number of nodes, whereas the traveled distance is the same in all cases. This input uncertainty naturally occurs in most real-life routing problems, especially in metropolitan areas where actual travel times between different points are almost impossible to predict. A solution for the time-dependent VRP with time windows was already proposed by Figliozzi (2012), although the work lacks of real time implementations as well as alternative route constructions.

In general, one of the main issues related to routing problems applied in an urban context with uncertainty related to the transportation costs is how to define realistic instances (Tadei, Perboli and Perfetti, 2017). Usually algorithms are compared bet-

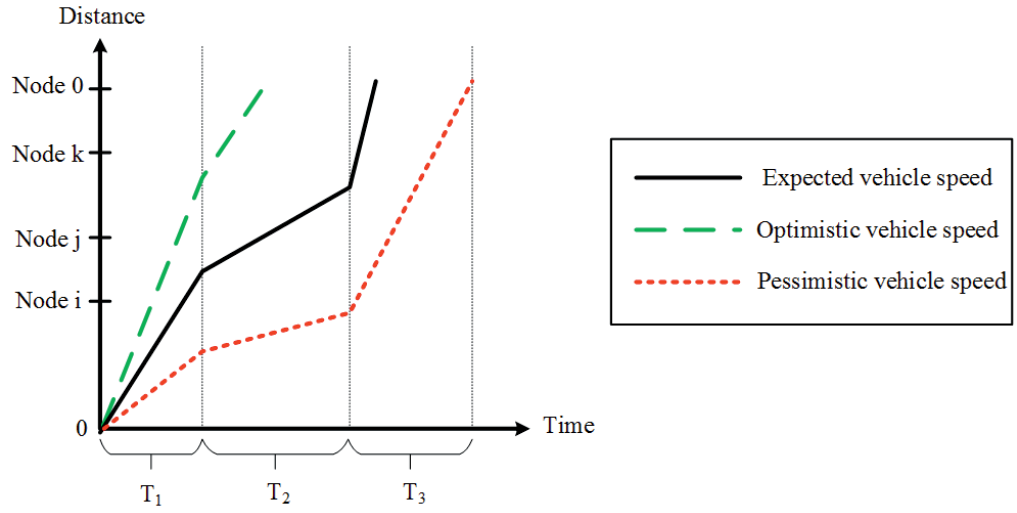


Figure 1: *The effect of stochastic travel duration due to time-varying vehicle speeds in time-dependent routing scenarios.*

ween them by using a common database. Although this is very helpful to compare algorithms, it is mandatory to connect the algorithm with real data from users to apply them.

This paper presents a simheuristic approach (Juan et al., 2018) to solve the time-dependent WCP with stochastic travel times (TDWCPST). By integrating Monte Carlo simulation into a metaheuristic framework, both time dependencies and stochastic travel speeds can be accounted for. Our metaheuristic framework combines biased randomization techniques (Quintero-Araujo et al., 2017) with an iterated local search algorithm (ILS) by Lourenço, Martin and Stützle (2003). The inclusion of a simulation procedure during the optimization process leads to a couple of advantages over traditional metaheuristic solving approaches. Apart from the consideration of stochastic travel times, it allows for a closer statistical risk analysis of the obtained solutions. This enables the creation of additional decision-making dimensions related to route robustness in uncertainty scenarios, e.g.: standard deviations or different quartiles obtained during the simulation phase. The implementation and performance of the solving methodology is tested on a large-scale case study. This case study refers to the waste collection process in the medium-sized city of Sabadell, which is located within the autonomous region of Catalonia, in northern Spain. It is important to note that real data from the waste collector department of the city is transformed to create real instances where to apply the algorithm.

Thus, the contributions of this work are threefold: (i) motivated by a real-life case, a rich TDWCPST is proposed; (ii) a large real-life data set, with several realistic routing constraints, is used to show the applicability of the proposed optimization procedure; and (iii) the potential of the simheuristic approach is illustrated in a range of computational experiments, hence yielding various managerial insights.

The paper is structured as follows: relevant literature on metaheuristic approaches for time-dependent routing problems and waste collection is reviewed in Section 2; the TDWCPST and the real-life problem setting are detailed in Section 3; Section 4 outlines our simheuristic solving framework; Section 5 describes different computational experiments and analyses obtained results; finally, Section 6 concludes this work and discusses possible future research directions.

2 Literature Review

This section reviews recent literature regarding metaheuristic solving frameworks for time-dependent VRPs and the WCP. For a more detailed overview on previous research regarding time-dependent routing problems the reader is referred to the work of Gendreau et al. (2015). A more extensive literature review on operational challenges and optimization methodologies in waste management is provided by Beliën, De Boeck and Van Ackere (2014) and Han and Ponce-Cueto (2015).

2.1 Metaheuristic solving methodologies for time-dependent routing problems

In the field of vehicle routing optimization, time dependency was not considered up to the early 2000s apart from a few exceptions. Malandraki and Daskin (1992) formulated travel times as a step function of the time of the day. This approach has the major drawback that the no-passing, first-in-first-out (FIFO) property is not guaranteed. Thus, a vehicle leaving node i might arrive later at node j than a vehicle leaving node i at a posterior starting time due to varying travel times. This drawback in the travel time function was improved by Hill and Benton (1992), who developed the first travel time model based on time-varying vehicle speeds, which implies the FIFO characteristic. Later, Ichoua, Gendreau and Potvin (2003) used an improved version of this vehicle speed model in combination with a parallel tabu search heuristic to show the benefits of time-dependent vehicle routing compared to its static counterpart. The impact of time-dependent travel times to avoid traffic congestion was also studied by Kok, Hans and Schutten (2012), who showed that late arrivals at customers and extra duty times through traffic jams can be significantly reduced through smart congestion-avoidance strategies.

An iterated local search algorithm for the time-dependent VRP with time windows (TDVRPTW) was presented by Hashimoto, Yagiura and Ibaraki (2008). Computational experiments include a variety of problem instances with up to 1,000 nodes. The TDVRPTW was also addressed in the works of Balseiro, Loiseau and Ramonet (2011) and Harwood, Mumford and Eglese (2013). The former developed an ant colony system hybridized with insertion heuristics which is tested on problem instances with up to 100 clients. The latter established quick estimates of time-dependent travel times for

the traveling salesman problem. Their results show that their estimations can lead to significant reductions in computation time. The TDVRP with simultaneous pickup and deliveries was addressed by Zhang, Chaovalitwongse and Zhang (2014) through an integrated ant colony and tabu search approach. A total of 100 customers were considered in their work. During the last decade, much attention has also been paid to the environmental effects of routing, in the context of the so called pollution routing problem. Kuo (2010) developed a simulated annealing algorithm for establishing emission minimizing vehicle routes while taking into account varying edge traversing times. Computational results are provided using benchmark instances with up to 100 customers. The trade-off between travel times and CO_2 emissions in time-dependent VRPs was analyzed by Jabali, Van Woensel and de Kok (2012). The time-dependent pollution routing problem was also analyzed in the work of Franceschetti et al. (2013). The authors proposed an integer linear programming formulation for cases without any traffic congestion. Environmental considerations are also included in the work of Soysal, Bloemhof-Ruwaard and Bektas (2015), who addressed the time-dependent two-echelon VRP through a comprehensive mixed integer linear programming (MILP) formulation.

All previously cited works focused on the deterministic version of the TDVRP. For stochastic problem settings the literature is more scarce. Lecluyse, VanWoensel and Peremans (2009) developed a tabu search metaheuristic for the TDVRP with stochastic travel times. Nahum and Hadas (2009) developed an extended version of the well-known savings algorithm to address the stochastic TDVRP. Tas et al. (2014) proposed a tabu search and adaptive large neighbourhood search metaheuristic for the TDVRP with soft time windows and stochastic travel times.

2.2 Metaheuristic frameworks in the optimization of waste collection

Different metaheuristic approaches have been presented in the solution of various WCPs and their extensions. Even though many works include a case study to show the real-life potentials of their frameworks, to the best of our knowledge, time dependency in the WCP has not yet been considered in the literature.

Baptista, Oliveira and Zúquete (2002) elaborated an extension of the Christofides and Beasley heuristic for the multi-period WCP modeled as a periodic VRP (PVRP) to combine vehicle scheduling over multiple time periods with route planning. The authors used their approach to improve municipal waste collection in the Portuguese city of Almeda. Also addressing a multi-period WCP, Teixeira, Antunes and de Sousa (2004) developed a cluster-first route-second heuristic to schedule and plan waste collection routes for different waste types in a case study in Portugal with over 1600 collection sites. Nuortio et al. (2006) presented a guided variable thresholding metaheuristic to solve a multi-period WCP with several thousand collection points in Eastern Finland. Hemmelmayr et al. (2013) addressed the PVRP with different waste types and up to 288 containers, which they solved with a variable neighbourhood search metaheuristic. They consider the landfills as intermediate facilities, which are inserted in pre-constructed

routes using dynamic programming. In the same work, the authors also discussed the single period WCP with multiple depots, in which the landfills serve as vehicle depots and disposal sites at the same time. Ramos, Gomes and Barbosa-Póvoa (2014) extended the typical objective of minimizing routing costs in order to include environmental concerns, considering multiple waste types and numerous vehicle depots in a case study in Portugal.

Only focusing on waste collection routing, Kim et al. (2006) developed an extension of Solomon's insertion algorithm to optimize routes of a North American waste management service provider, considering a capacitated vehicle fleet, time windows, and driver lunch breaks. The authors reported reduced routing distances of up to 10%. Furthermore, a benchmark set of 10 realistic instances based on the original case study ranging from 102 TO 2,100 nodes is provided. Using the same benchmark set, Benjamin and Beasley (2010) combined tabu search with a variable neighbourhood search metaheuristic. By exchanging containers and landfills within and between routes, the solution search space is systematically increased. Likewise, Buhrkal et al. (2012) put forward an adaptive large neighbourhood search metaheuristic. Based on an initial solution, their approach applies a range of destroy-and-repair methods to examine several solution neighbourhoods. It is called adaptive since the choice of methods depends on the solution quality obtained during the construction of earlier solutions. Moreover, an acceptance criterion for new solutions based on simulated annealing is included. Likewise, Markov, Varone and Bierlaire (2016) presented a multiple neighbourhood search heuristic for a real-word application of the waste collection VRP with intermediate facilities. The authors consider a heterogeneous vehicle fleet and flexible depot destinations in their approach. Gruler et al. (2017a) developed a metaheuristic algorithm to assess the potentials of horizontal collaboration in urban waste collection.

Concerning the WCP under input uncertainty, the literature is more scarce with most works focusing on stochasticity concerning expected waste levels. Ant colony optimization and a hybrid approach based on a genetic algorithm and tabu search for a case study with 50 containers in Malaysia is presented in Ismail and Irhamah (2008) and Ismail and Loh (2009). After planning *aprioristic* routes, waste levels are simulated according to a discrete probability distribution. Routes undergo a recourse action (i.e., an additional disposal trip) whenever actual demand exceeds the planned collection amount. Nolz, Absi and Feillet (2014) formulated a collector-managed inventory routing problem for a case study on the collection of infectious waste. By using real information obtained through radio frequency identification, their adaptive large neighbourhood search algorithm is able to consider stochastic waste collection levels. Alshraideh and Abu Qdais (2017) combined a multi-period WCP with time windows and stochastic demands in a real case study of medical waste collection from 19 hospitals in Northern Jordan. They used a genetic algorithm and a probability constraint regarding a pre-defined service level to solve the problem. Also, Gruler et al. (2017b) presented a variable neighbourhood search based simulation-optimization approach for the WCP with stochastic demands. Although metaheuristics are becoming the predominant methodology in solving

WCP under rich and realistic scenarios (Hannan et al., 2018; Asefi et al., 2019), other approaches such as mixed-integer programming are also being employed by some experts (Mohsenizadeh, Tural and Kentel, 2020).

3 Problem Description

This section outlines the real-life case study of collecting waste under different routing constraints and travel time assumptions in the city of Sabadell. Furthermore, the time-dependent WCP with stochastic travel times (TDWCPST) and the applied travel speed model for different time periods is discussed in more detail.

3.1 The waste collection problem in Sabadell

Sabadell is a medium-sized city of roughly 200,000 inhabitants located within the autonomous Spanish region of Catalonia. Collection vehicles are located at a central depot and collected garbage is disposed in a single landfill. Expected waste levels in each container, average service times at each node, and the average vehicle travel speeds during different time periods are known. The problem settings consists of a total of 921 paper waste containers which are currently visited on 9 different routes. The locations of the vehicle depot, the landfill, waste containers, and the original route assignation can be seen in Figure 2 (the central depot and the landfill are marked by the square symbols).

According to the managers, in a scenario with dynamic travel times as the one being considered, the total time required to complete the waste collection process is the main key performance indicator. On the one hand, the operational times directly affect the operational costs associated with the waste collection process in terms of wages and vehicle usage costs. On the other hand, an important routing constraint is that collection routes need to be completed between 9 a.m. and 4 p.m., as these are the opening hours of the central depot at which the collection vehicles are stationed. Moreover, different time periods within the daily planning horizon can be identified regarding expected traffic speeds:

- Heavy traffic on all streets is expected during the rush hour from 9 a.m. to 10 a.m. and from 1 p.m. to 2 p.m.
- Traffic jams are expected in streets close to primary schools in the time periods of 9 a.m. to 10 a.m., 12 p.m. to 1 p.m., and 3 p.m. to 4 p.m.

Especially the latter observation is of importance in the planning of waste collection routes. Containers in the affected streets should not be visited within the depicted time period. Due to parents picking up their children from primary schools, streets within a certain distance radius of the school building should be avoided in the given period if

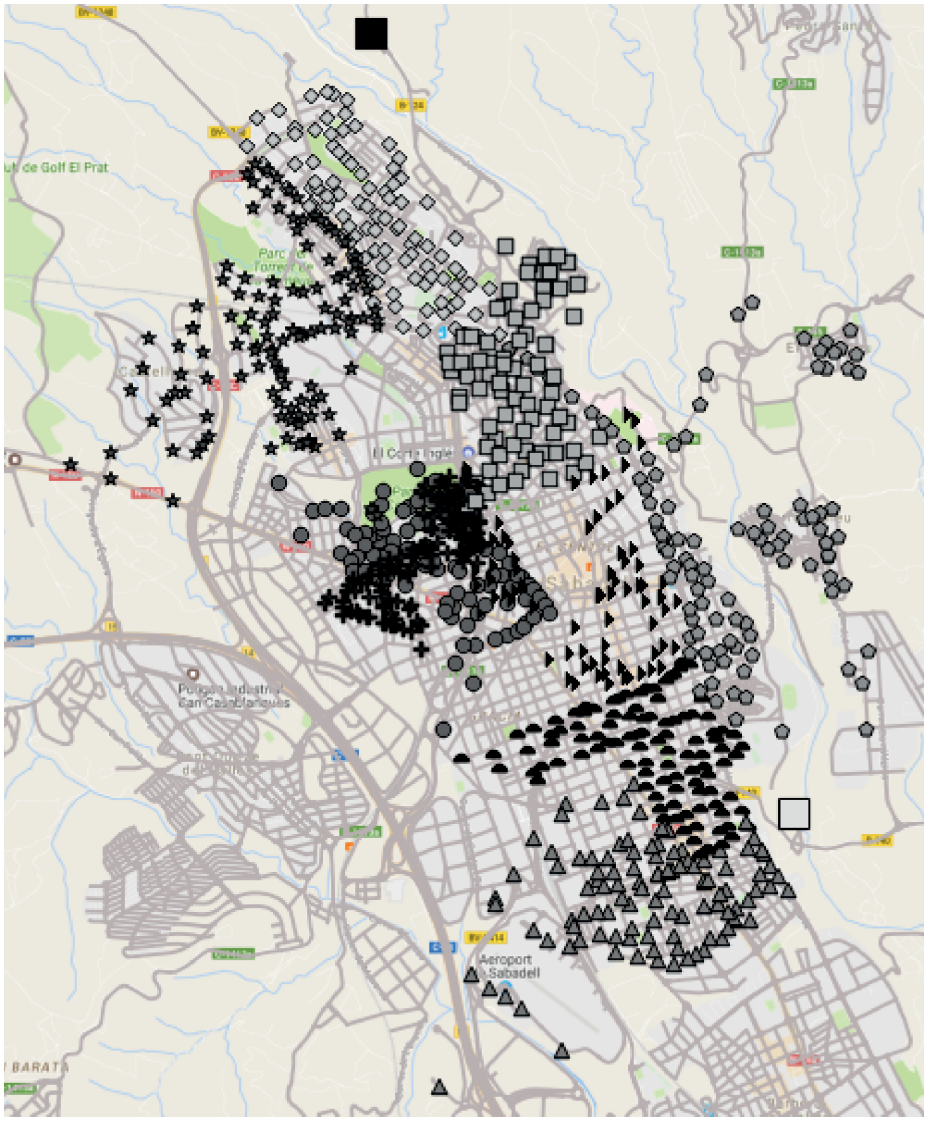


Figure 2: *Node locations and original route assignation.*

possible. According to the experience of the decision-taker, a radius of 500 m around primary schools is considered. Apart from delays in the collection process, visiting these streets during the most busy hours affects many citizens and can even be dangerous due to children exiting the primary school facilities. The influenced streets in the city centre of Sabadell for which the additional constraints apply are highlighted in Figure 3.

3.2 A time dependent travel speed model for the WCP

The TDWCPST can be described on a graph $G = (V, E)$:

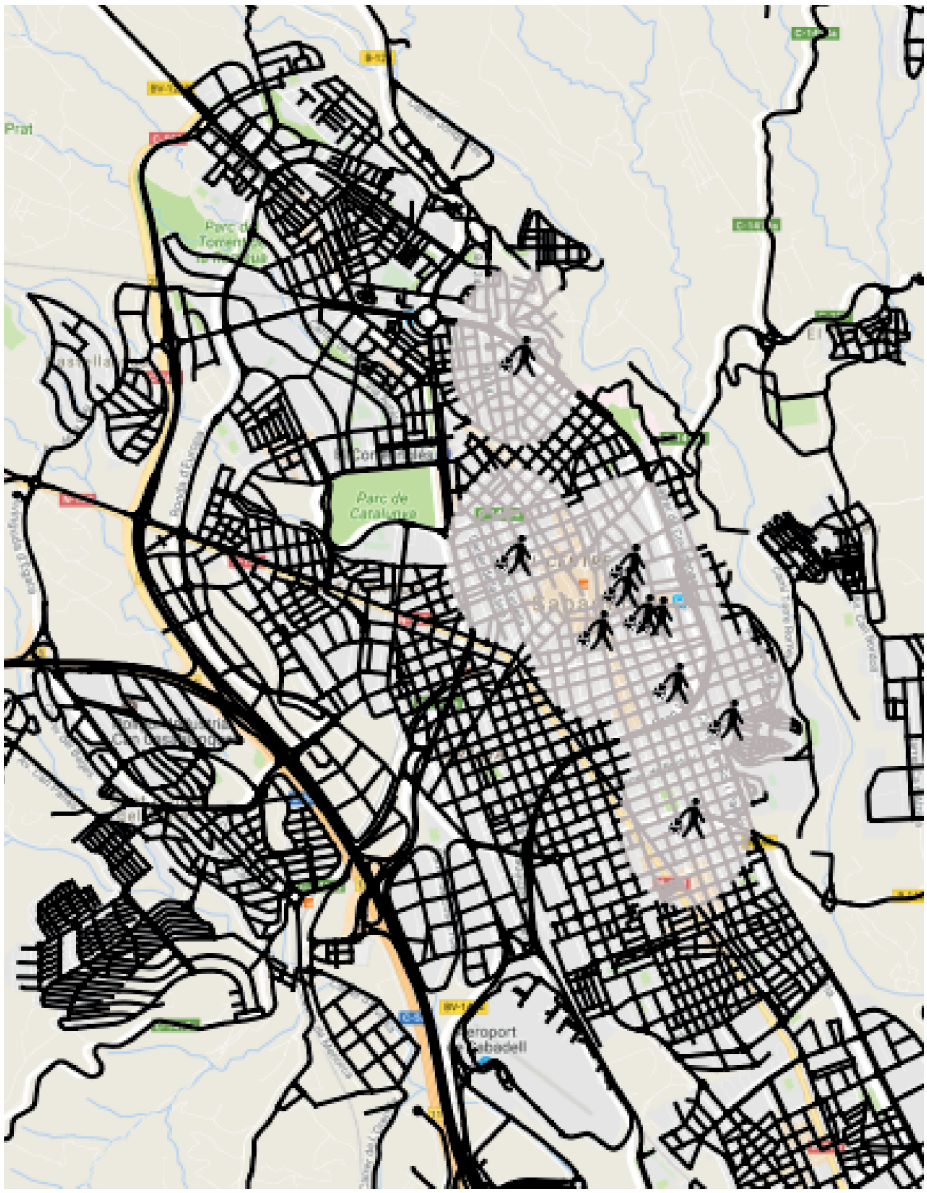


Figure 3: *Streets to be avoided during highly occupied traffic periods.*

- Node set $V = V^d \cup V^f \cup V^c$ includes:
  - *(i)* A central depot $V^d = \{0\}$ at which a homogeneous fleet of waste collection vehicles, each of them with capacity C, is located.
  - *(ii)* A set $V^f = \{1, 2, \ldots, m\}$ of m landfills at which collected waste must be disposed if vehicle capacities are reached and before a vehicle returns to the central depot (making more than one landfill trip per route possible if no other route constraints are violated).

(iii) A set of n waste containers $V^c = \{m+1,\ldots,m+n\}$ with associated waste levels $q_i > 0$ $(\forall i \in V^c)$. Service times for emptying any container and for disposing collected garbage at any landfill are defined as $s_i > 0$ $(\forall i \in V \setminus V^d)$.

- Edge set $E = \{(i,j)/i,j \in V, i \neq j\}$ describes all edges connecting any two nodes.
- Travel distances $d_{ij} \geq 0$ between any two nodes in V are known.
- Additional routing constraints include a maximum amount of waste to be collected during each route and the maximum route duration defined by the opening and closing times at the central depot.

Our travel speed model for time varying vehicle velocities is based on the discussions of Ichoua et al. (2003). The planning horizon (defined by the depot opening hours) is divided into p time periods $T_1, T_2, \ldots, T_p$. Travel durations tt_{eT} to cross any edge in $e \in E$ can be calculated as the quotient of travel distances and vehicle speeds v_T $(T \in \{T_1, T_2, \ldots, T_p\})$, such that $tt_{eT} = d_e/v_T$. In the specific case of waste collection in Sabadell, different travel speeds can be defined for different edges, e.g., due to rush hour traffic or other events such as opening or closing hours of schools. For this reason, edge set E is partitioned into S subsets with E_s $(s = 1,2,\ldots,S)$. Thus, travel speeds can be formulated as tt_{sT} to show the travel speed of any edge of the edge subset E_s during time period T. This step-wise travel speed model along different times of the planning horizon is a natural way of estimating travel duration of different edges in real-world conditions. Furthermore, it implies the satisfaction of the FIFO property.

4 Solving Framework

The different stages of the proposed simheuristic solving methodology for the TD-WCPST are summarized in Figure 4. By integrating simulation into a biased-randomized iterated local search (BR-ILS) algorithm, a set of promising stochastic solutions are constructed. These solutions are then refined in a more intensive simulation procedure. Finally, the defined set of solutions undergoes a more detailed risk analysis according to different criteria. All steps are outlined in more detail in the following subsections.

4.1 Constructing an initial time-dependent WCP solution

Our approach starts by constructing a feasible initial solution with an enhanced framework of the well-known savings heuristic for routing problems (Clarke and Wright, 1964). In the original procedure, the savings s_{ij} of including any edge e connecting two customers i and j in a constructed solution are calculated as $s_{ij} = s_{ji} = c_{i0} + s_{0j} - c_{ij}$. However, this assumption does not hold in the special case of waste collection, as the round trip costs between the central depot and any waste container are asymmetric due to the additional landfill visit at the end of any completed route. Thus, the route travel

direction influences the savings values assigned to each edge. In order to account for the necessary landfill visit in every route, the expected savings of each edge are calculated as average values of completing a route in both directions, such that $E[s_{ij}] = (s_{ij} + s_{ji})/2$. After each merge, this initial estimate is updated to account for the real travel times depending on the hour of the day.

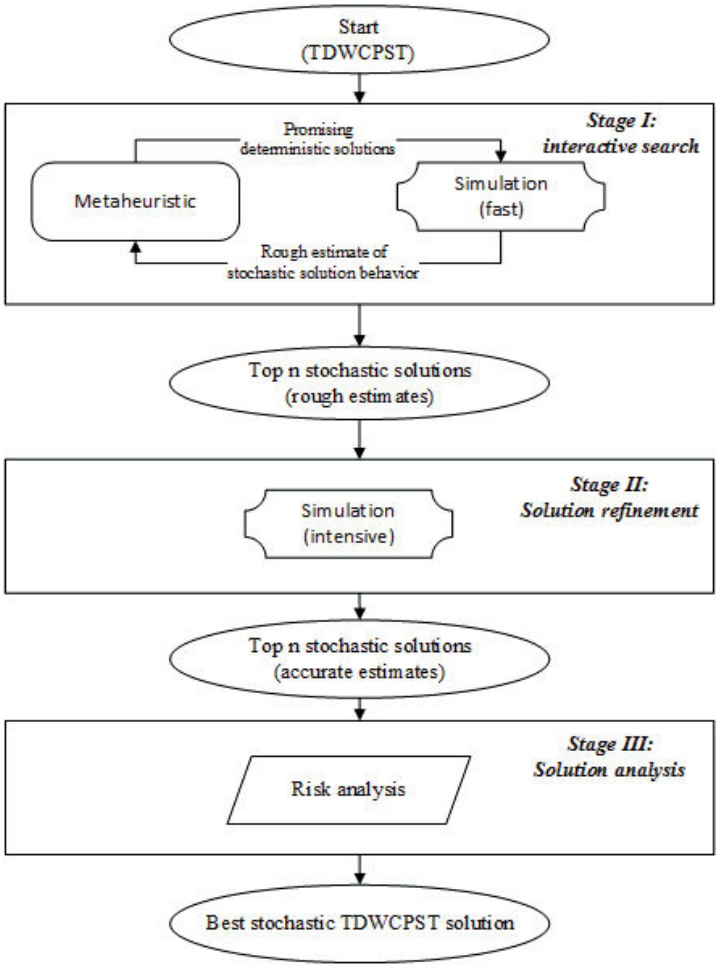


Figure 4: *Simheuristic solving methodology.*

Apart from this algorithm adaption to the problem setting, we enhance the greedy edge selection process of the savings procedure through a probabilistic construction behaviour based on biased randomization techniques (Ferone et al., 2019). As highlighted in Algorithm 1, a candidate set of edges is ranked according to their respective savings value. In the following, a feasible waste collection route is created by iteratively adding solution elements from the eligible edges. Selection probabilities follow a geometric distribution defined through parameter α ($0 < \alpha < 1$), which depicts the probability of the most promising solution element to be chosen. This process is similar to

the GRASP procedure discussed in the work of Resende and Ribeiro (2010). However, while GRASP is based on a restricted candidate list and a uniform selection probability, selection probabilities are inclined to more promising solution elements –which are all potentially eligible at each solution construction step– in this biased randomization approach. In a biased-randomized algorithm, the choice of the skewed probability distribution has an impact on the quality of the final solution. As discussed in Grasas et al. (2017), the geometric and the decreasing triangular probability distributions have been successfully used in previous work, but other probability distributions (either theoretical or empirical) are possible as well.

Algorithm 1: Biased randomization to create an initial TDWCP solution

Input: Skewed probability distribution f, parameter α, edge set E
$sol \leftarrow \emptyset$
initialize candidate set: $CL \leftarrow E$
sort CL according to savings value
while *solution sol is not complete* **do**
  Randomly select $pos \in CL$ according to distribution function $f(\alpha)$
  $sol \leftarrow sol \cup pos$
  $CL \leftarrow CL \setminus pos$
  sort CL
end
return TDWCP solution *sol*

4.2 A simheuristic framework for the time-dependent WCP with stochastic travel times

Our simheuristic procedure to solve the TDWCPST is outlined in Algorithm 2. Once an initial solution is constructed and set as the current incumbent *baseSol* and *bestSol* solutions, an iterated local search algorithm is started (Lourenço et al., 2003). During a predefined stopping criterion, new TDWCP solution neighbourhoods are created by perturbating the current *baseSol*. Each perturbated solution *newSol* then undergoes a local search phase to find the local minimum within the current solution structure. As perturbation operator a double-bridge move is applied. Hereby, a solution is partitioned into four pieces of random size, which are subsequently joined in an arbitrary order. As local search movement, a *2-opt* operator is employed (Muyldermans et al., 2005).

Up to this point, deterministic (expected) travel duration between difference network nodes are considered. In order to account for uncertainty in input variables, Monte Carlo simulation is applied to any promising solution found in the metaheuristic search. A TDWCP solution *newSol* is deemed promising if its deterministic travel duration outperform those of the currently incumbent *baseSol* or if if a simulated annealing-like acceptance criterion is met. The travel duration between all edges of a promising solution are simulated from a log-normal probability distribution during *nSim* simulation runs. At this stage any other probability distribution could be applied, but the log-normal one is a "natural" choice to model non-negative random variables, such as travel times

in routing problems or times-to-failure in reliability studies (Faulin et al., 2008). During each simulation iteration, expected travel duration tt_{sT} between any two points are defined as distribution mean of the probability function. Variance factor k defines travel duration variance levels. With $E[tt_{sT}] = tt_{sT}$ and $Var[tt_{sT}] = k \cdot tt_{sT}$, the location parameter μ_i and scale parameter σ_i defined for the probability function can be formulated as:

$$\mu_i = \ln(E[tt_{sT}]) - \frac{1}{2} \cdot \ln\left(1 + \frac{Var[tt_{sT}]}{E[tt_{sT}]^2}\right)$$

$$\sigma_i = \left|\sqrt{\ln\left(1 + \frac{Var[tt_{sT}]}{E[tt_{sT}]^2}\right)}\right|$$

As a result of time varying travel speeds, the variability in solution waste collection durations estimated after each simulation run can be expected to increase with higher variance levels. In particular, waste collection close to primary school locations is penalized by significantly reduced travel speeds during predefined time periods. After the simulation phase, the stochastic travel durations of *newSol* are defined as the average of all simulation results.

If the stochastic costs of the considered solution outperform the estimated stochastic travel durations of the incumbent *baseSol* and/or *bestSol*, they are updated respectively. Moreover, each solution that is defined as incumbent *baseSol* during any stage of the simheuristic procedure is included in a TDWCPST solution set $eliteSols$. After the algorithm stopping criterion is reached, solutions included in this exclusive set of elite solutions undergo a more intensive simulation phase defined by a higher number of simulation runs. This allows a more accurate estimation of the best found WCP solutions in stochastic travel time scenarios.

The described combination of simulation with metaheuristics leads to several advantages over deterministically focused optimization approaches. Firstly, the search phase is driven by the stochastic solution estimates obtained during the simulation (i.e., the *baseSol* is updated according to the cost estimates provided by the simulation component). Secondly, TDWCPST solutions can be realistically evaluated under different uncertainty scenarios. Finally, the simheuristic methodology allows the evaluation of different solutions according to additional criteria instead of simply focusing on the defined objective function. Due to the stochasticity in real-life travel duration, the completion of waste collection plans is likely to vary with respect to the predicted driving times. For this reason, decision-makers need a more insightful decision support than simply focusing on the minimization of expected travel times. Thus, we implement a final risk analysis for the elite solutions in our simheuristic procedure. At this stage, additional dimensions related to the robustness of a considered solution, such as the standard deviation or the quartiles, are computed.

Algorithm 2: A simheuristic for the TDWCPST

```
   Input: f, E, α, nSim_short, nSim_long, k
 1 nodes ← getNodes(E)
 2 costMatrix ← getCostMatrix(E)
 3 initSol ← generateBRSolution(f, α, E)                    // Biased-Randomized Algorithm
 4 baseSol ← initSol
 5 stochDuration(baseSol) ← infinite
 6 bestSol ← baseSol
 7 elitSols ← ∅
 8 while stopping criterion not reached do
 9 |   newSol ← perturbate(baseSol, costMatrix)              // perturbation stage
10 |   newSol ← localSearch(newSol, costMatrix)             // local search stage
11 |   delta ← detDuration(baseSol) − detDuration(newSol)
12 |   if delta ≥ 0 then
13 |   |   credit ← delta
14 |   |   stochDuration(newSol) ← simulation(newSol, nSim_short, k)
15 |   |   if stochDuration(newSol) ≤ stochDuration(baseSol) then
16 |   |   |   includeInEliteSolutionSet(newSol)
17 |   |   |   baseSol ← newSol                            // simulation driven baseSol
18 |   |   |   if stochDuration(newSol) < stochDuration(bestSol) then
19 |   |   |   |   bestSol ← newSol
   |   |   |   end
   |   |   end
   |   end
20 |   else if −delta ≤ credit then
21 |   |   credit ← 0
22 |   |   stochDuration(newSol) ← simulation(newSol, nSim_short, k)
23 |   |   baseSol ← newSol
   |   end
   end
24 for eliteSol ∈ eliteSols do
25 |   stochDuration(eliteSol) ← simulation(eliteSol, nSim_long, k)
   end
26 return bestSol
```

4.3 Creating a real-life distance matrix

In this subsection we show the process followed to generate a real-life distance matrix. The data was obtained through a collaboration agreement between the Internet Computing and Systems Optimization (ICSO@IN3) research group and the company SMATSA, which is responsible for the collection of waste in the inner-city area of Sabadell. The problem dealt in the present paper involves the waste disposal vehicle routing in order to design efficient routes between 886 paper waste containers. A single depot and landfill are considered. Locations are given as Longitude/Latitude (Long/Lat) and postal addresses are also available. The goal is to create a real-life distance matrix from these data.

In the creation of this distance matrix only open software has been used. In particular, we have used: *QGIS 2.18* (`https://www.qgis.org`) as a geographic information system (GIS); *PostGIS* (`https://postgis.net`), which is the geographic extension of the database *PostgreSQL* (`https://www.postgresql.org`); and *pgRouting* (`https://pgrouting.org`) to obtain the distances between pairs of locations. In addition, *Open Street Map* or OSM (`https://www.openstreetmap.org`) has been em-

ployed as a base map. The first step is to download the base map for the zone of Sabadell from OSM. This downloaded file is then processed with *osm2po* (`https://osm2po.de`) to transform it into a routable file. One of the outputs of the program is an SQL file that can be executed in *PostgreSQL* with the *PostGIS* extension. The output is a table that can be visualized in *QGIS* with the *Add PostGIS table* function.

At this step the map is available in a GIS and, although the aspect is visually correct, it does not have topology, which is required to connect nodes and to obtain real distances. The topology can be created with the *pgrouting* query *pgr_createTopology*, which can be run from *QGIS* thanks to the database manager plug in. Once the map has a topology, *pgRouting* can be run to obtain the shortest path between two given points. Figure 5 shows the uploaded map with the route between two points.

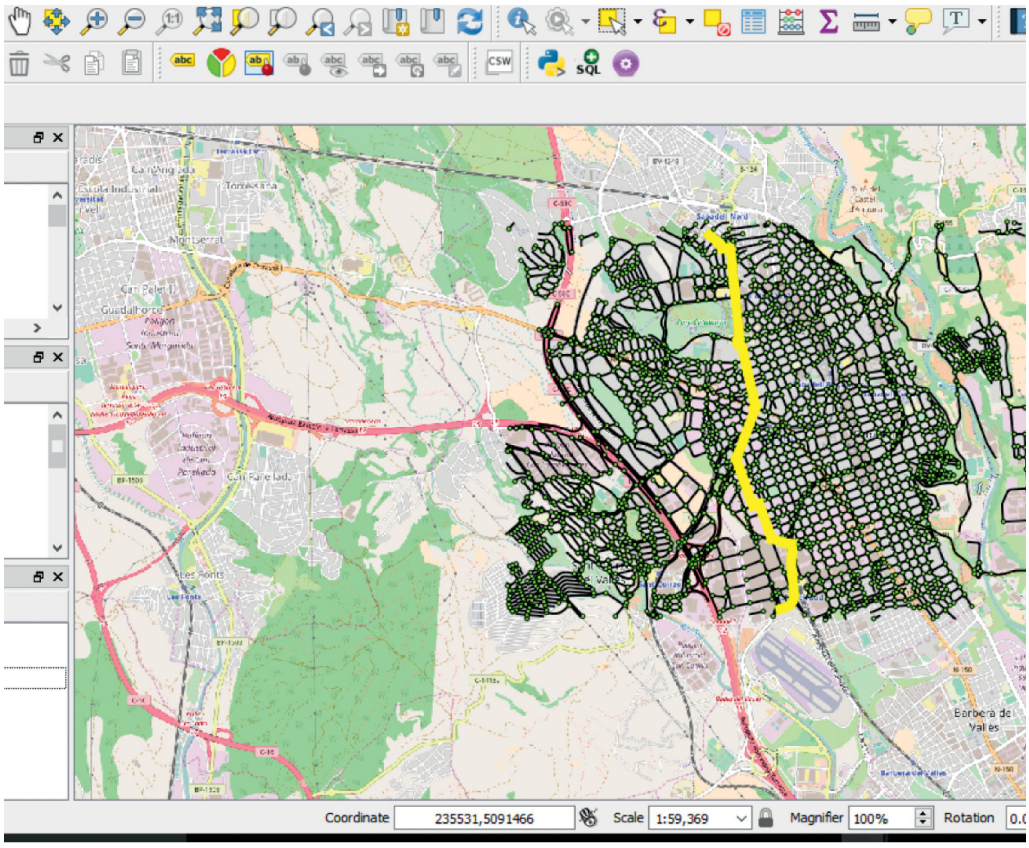


Figure 5: *Route between two points obtained with* pgRouting *in* QGIS *after uploading the base map and having introduced the topology.*

After this process, we had the map of the working area in a GIS and prepared to obtain the distance between any two points. The next step is uploading the location of the 886 containers, the depot and the landfill. Although the location is quite precise, there

may be some errors and it is important to verify out-layer nodes. Usually these points can be visually located in the map and corrected using the postal address (Figure 6).

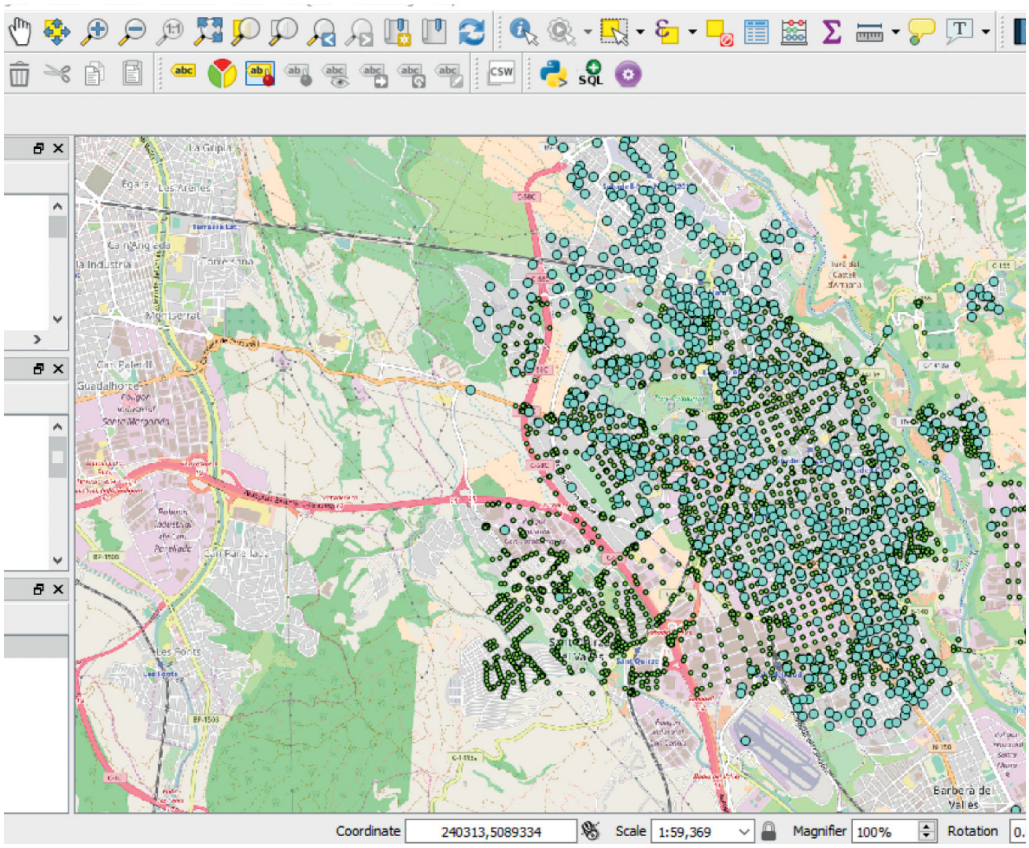


Figure 6: *Network nodes and imported point location layer.*

On the other hand, the position of the nodes does not correspond to the street lines of the map in QGIS, since the streets have a width and they are represented as a line axe. Therefore, every point has to be linked to its corresponding street axis. It has been done with the *NNJoin QGIS* plug in, which obtains the nearest neighbour between every single point and the start / end node of street axes. Figure 6 shows the nodes within the network and the imported nodes. Thus, it has been possible to obtain the closest point from the axes to every single container, which generates a new list of points, but this time, within the routable network. This introduces a little error in the position of the containers if they are not in the end node of the street axe. A more precise solution would have been to locate the nearest point of the axe and split the axe at that point. However, since containers are usually located at the corner of the streets, and according to the managers' opinion, the error introduced can be considered as a non-relevant one for the purposes of this study.

Finally, we can obtain the distance matrix using *PostgreSQL* (with the extensions *PostGIS* and *pgRouting*). The distance in km is set as cost, and a distance matrix of the first 10 nodes is established. The name of the output file is *sabadel_2po_4pgr*.

5 Computational Experiments and Analysis of Results

The proposed simheuristic solving framework is applied to the real-life waste collection problem setting described in section 3. The algorithm is implemented as a Java application and tests are run on a personal computer with 4GB RAM and an Intel Pentium®processor with 2.16GHz. The necessary algorithm parameters to complete the described tests are specified as follows. These parameters have been obtained after a quick calibration based on the methodology proposed by Calvet et al. (2016):

- Skewed probability distribution f: geometric with parameter α = 0.3.
- $nSim_{short}$: 100.
- $nSim_{long}$: 1000.
- BR-ILS stopping criterion per instance: 30 seconds.

According to the observations of the decision maker, the average vehicle speed in normal traffic conditions is 25 km/h. The travel speed is divided by 5 and 25 during heavy traffic and traffic jams, respectively. Average vehicle service times at each container are set to 90 seconds, while 45 minutes are necessary to empty a vehicle at the landfill. Stochastic travel times are generated with three different variance factors, $k = 1, 2.5, 10$, representing different (low / medium / high) uncertainty levels. All variance scenarios are represented in Figure 7, showing the travel times of edge subset s during time period T with an expected traversing time of $E[tt_{sT}] = 25$ time units. The shadowed area under each curve represents 95% of the simulated values. In the low-variance scenario ($k = 1$), 95% of actual driving times fall between 16.64 and 36.15 time units with a high density around the expected value. As the variance level is increased, the maximum density of the simulated times for all edges decreases and a higher variability can be observed. Note that the overall driving duration of a solution will increase as the expected travel time uncertainty increases. Moreover, the special case of $k = 0$ is equivalent to the deterministic routing case.

5.1 Experimental results

In order to evaluate the performance of our simheuristic algorithm, its results are compared to the nine waste collection routes currently completed on a daily basis in Sabadell. The comparison of the current routes (i.e., those built by SMATSA) and the best found solution of the BR-ILS algorithm in different variance scenarios is listed in Tables 1

(deterministic case and low variance) and 2 (medium and high variance). A total of 10 independent executions were run (each one using a different seed for the random number generator), and the best-found solution was returned by the algorithm. Each route holds between 79 and 122 waste containers to be emptied. To allow a fair comparison, the current order of visiting waste containers is evaluated in accordance with the necessary algorithm parameters described before. Since the distance matrix has been created using real-life distances, it is possible to make the comparison in order to know if the cost function is actually improved.

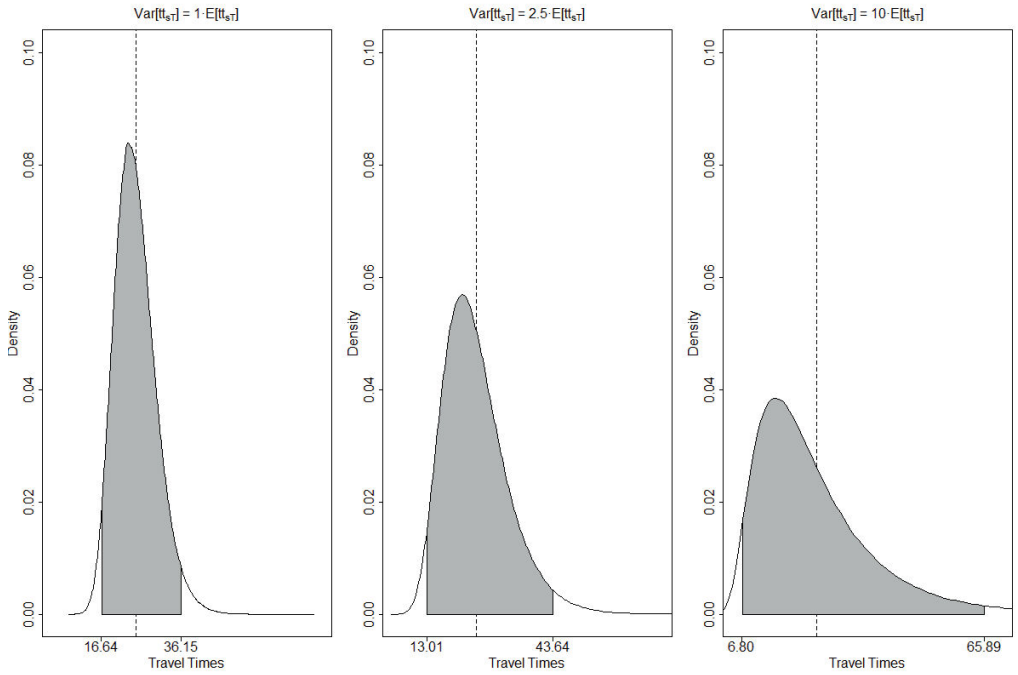


Figure 7: *Log-normal distribution of different variance levels around an expected travel time of* 25 *time units.*

It is important to note that comparing the nine routes separately, instead of designing new routes, allows us to compare our algorithm with the current situation in a real problem. The comparison is performed in terms of total time employed in completing the collection process, since this is the main key performance indicator for the managers.

In all travel duration variance scenarios, the BR-ILS is able to significantly outperform the current waste collection routes (by over 12% on average). Moreover, the solution travel duration in different uncertainty scenarios provided by our metaheuristic show that estimated travel duration increase with higher variance levels. The best results with the simheuristic BR-ILS algorithm are obtained when considering all 921 waste containers in a "global" waste collection instance. In this case, new route-to-containers assignments are established instead of solely focusing on reordering pre-established waste collection routes. For example, in the deterministic routing case, and

using a running time of 120 seconds, the global solution yields a overall driving duration of 2,820.5 minutes, with only 8 necessary garbage collection routes, thus saving one route to the company. Regarding the solution to the stochastic version of the problem, the proposed simheuristic has been run for a maximum time of 5 minutes before returning the best-found solution. This maximum computational time was suggested by the managers, who have to plan the collection routes every morning.

Table 1: *Driving duration in minutes of current routes compared to our best found solution (deterministic case and low variance scenario).*

Route	**k = 0**		**Diff (%)**	**k = 1**		**Diff (%)**
	Current [1]	**Our Best [2]**	**([2]-[1])/[1]**	**Current [3]**	**Our Best [4]**	**([4]-[3])/[3]**
1	392.5	357.3	−9.0	391.3	362.7	−7.3
2	379.9	265.0	−30.2	381.3	271.6	−28.8
3	470.8	345.5	−26.6	455.5	346.0	−24.0
4	386.4	342.3	−11.4	384.2	341.8	−11.0
5	374.3	335.1	−10.5	387.9	342.7	−11.7
6	396.2	371.4	−6.2	397.1	388.7	−2.1
7	372.8	340.4	−8.7	364.2	340.6	−6.5
8	393.9	326.7	−17.0	399.8	342.1	−14.4
9	407.9	323.0	−20.8	411.1	323.6	−21.3
Total	***3,574.5***	***3,006.7***	***–15.9***	***3,572.3***	***3,059.9***	***–14.3***

Table 2: *Driving duration in minutes of current routes compared to our best found solution (medium and high variance scenario).*

Route	**k = 2.5**		**Diff (%)**	**k = 10**		**Diff (%)**
	Current [1]	**Our Best [2]**	**([2]-[1])/[1]**	**Current [3]**	**Our Best [4]**	**([4]-[3])/[3]**
1	391.6	363.0	−7.3	392.7	367.5	−6.4
2	381.4	281.3	−26.2	388.1	293.4	−24.4
3	457.9	349.8	−23.6	460.2	363.3	−21.0
4	383.9	342.6	−10.7	384.2	340.2	−11.4
5	389.7	342.7	−12.0	395.4	355.2	−10.2
6	397.2	386.9	−2.6	397.3	387.5	−2.5
7	362.3	337.7	−6.8	360.4	340.1	−5.6
8	398.8	344.0	−13.8	398.9	355.0	−11.0
9	410.6	329.4	−19.8	414.0	346.5	−16.3
Total	***3,573.3***	***3,077.3***	***–13.9***	***3,591.1***	***3,148.7***	***–12.3***

5.2 Risk analysis of different TDWCPST solutions

For each waste collection plan (i.e., for each solution to the TDWCPSD), our simheuristic algorithm not only generates information about its expected travel duration and ex-

pected driving distance, but it can also provide the plan's probabilistic profile (including risk and reliability analyses). Thus, statistical values such as the standard deviation of travel duration, the median, or the third quartile can be obtained during the simulation runs without increasing the computing effort.

Table 3 shows different attributes of three elite solutions of the global TDWCPSD with all garbage containers in a high variance scenario ($k = 10$). The deterministic and stochastic travel duration, driving distance, standard deviation, median, third quartile, and the number of waste collection routes of each TDWCPST solution are provided. As highlighted in the radar chart shown in Figure 8, each solution outperforms the others in a different decision-making dimension. While solution B is the most promising solution regarding deterministic travel duration, solution A shows the best results in terms of expected travel times and overall travel distance. However, the standard deviation of travel duration obtained during the long simulation run (which can be seen as a reliability indicator of a given solution) is the lowest for solution C. This solution behaviour is also observed in the multiple boxplot shown in Figure 9. It can be clearly seen that the most promising deterministic solution B yields the highest travel duration variance, suggesting a low reliability of the constructed waste collection routes. Likewise, the median and third quartile could be considered in a closer risk analysis according to the preferences of the waste collection route planner. Since this work is addressed to a real situation scenario, it is important for the planner this degree of freedom that allows to find different solutions. In this experiment we offered three solutions to the planner or decision maker. In other settings, the specific number could be adjusted taking into account the magnitude of the differences among solutions and the preferences of the planner.

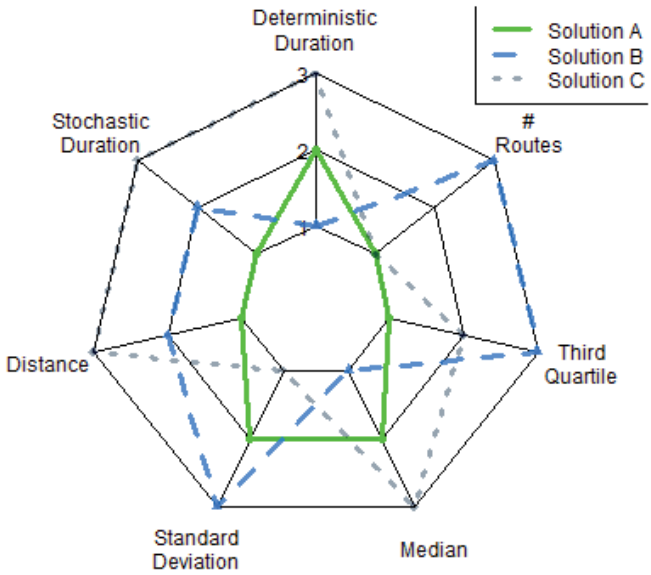


Figure 8: *Ranking of TDWCPST solutions according to different quality dimensions.*

Table 3: *Analysis of different TDWCPST solutions (high variance scenario).*

Solution	Det. Duration (min)	Stoch. Duration (min)	Distance (km)	Stand. Dev.	Median	Third Quartile	# Routes
A	2,879.61	2,936.05	264.58	31.98	2,932	2,956	8
B	2,827.49	2,938.67	278.45	64.59	2,930	2,969	9
C	2,897.58	2,952.18	280.27	26.46	2,951	2,967	8

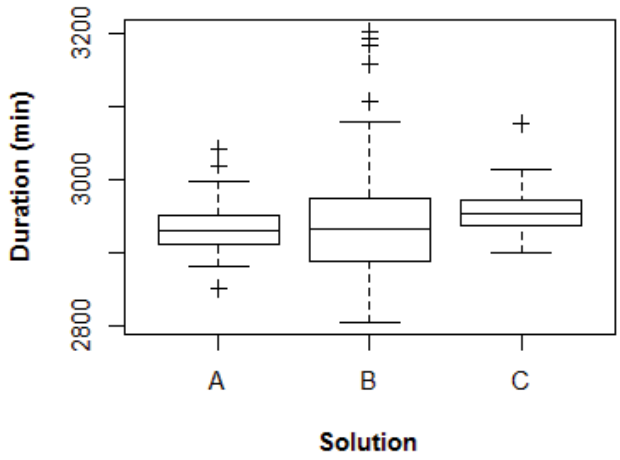


Figure 9: *Comparison of simulation results of different TDWCPST solutions.*

6 Conclusions

This work presents a simheuristic algorithm for the time-dependent waste collection problem with stochastic travel times to improve the real-life case of the waste collection process of several hundred waste containers in the Spanish city of Sabadell. The algorithm works by integrating simulation into a metaheuristic framework, which is based on a biased-randomized iterated local search. Uncertainty in travel duration between different nodes in a the city logistics network is considered as well.

The work also shows the process followed to obtain the real-life distances. Working with real-life distances allows the comparison of the algorithm results with the real routes that are used in Sabadell nowadays. Results suggest significant travel duration reductions in different variance scenarios. Furthermore, a risk analysis of obtained solutions along different dimensions such as the standard deviation of travel duration is performed. The results underline the importance of risk aware route planning in the process of waste collection.

The research completed in this paper can be extended in several directions. Although a simheuristic algorithm has been used to obtain garbage collection routes with real-life distances to compare then with the routes of Sabadell, other standard algorithms of the literature could also be tested and compared with them. This work would allow to see the relative difference between several algorithms in a real situation. Moreover, different procedures to generate the initial solution can be tested and their effect on the global performance of the algorithm can be assessed.

In addition, our simheuristic procedure could be extended to consider additional input variables that are typically shaped by some kind of stochastic behaviour, e.g.: waste to be collected or even waste containers themselves. Similarly, the problem setting could be enriched by including historical data to construct a more realistic travel speed model for the study area. An interesting concept in this context is the emerging technique of learnheuristics (Calvet et al., 2017), which complements the simheuristic solving framework by including machine learning techniques to consider problem dynamic inputs –e.g., varying traffic conditions at different times.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science (Spanish Ministry of Science (PID2019-111100RB-C21 / AEI / 10.13039 / 501100011033, RED2018-102642-T), AGAUR and FEDER (2018 LLAV 00017), and the Erasmus+ programme (2019-I-ES01-KA103-062602). Furthermore, we would like to thank the company SMATSA for providing us with their data and support.

References

Alshraideh, H. and Abu Qdais, H. (2017). Stochastic modeling and optimization of medical waste collection in Northern Jordan. *Journal of Material Cycles and Waste Management*, 19, 1–11.

Asefi, H., Lim, S., Maghrebi, M. and Shahparvari, S. (2019). Mathematical modelling and heuristic approaches to the location-routing problem of a cost-effective integrated solid waste management. *Annals of Operations Research*, 273, 75–110.

Balseiro, S. R., Loiseau, I. and Ramonet, J. (2011). An ant colony algorithm hybridized with insertion heuristics for the time dependent vehicle routing problem with time windows. *Compututers & Operations Research*, 38, 954–966.

Baptista, S., Oliveira, R. C. and Zúquete, E. (2002). A period vehicle routing case study. *European Journal of Operational Research*, 139, 220–229.

Beliën, J., De Boeck, L. and Van Ackere, J. (2014). Municipal solid waste collection and management problems: a literature review. *Transportation Science*, 48, 78–102.

Benjamin, A. M. and Beasley, J. E. (2010). Metaheuristics for the waste collection vehicle routing problem with time windows, driver rest period and multiple disposal facilities. *Computers & Operations Research*, 37, 2270–2280.

Bing, X., Bloemhof, J. M., Ramos, T. R. P., Barbosa-Povoa, A. P., Wong, C. Y. and van der Vorst, J. G. (2016). Research challenges in municipal solid waste logistics management. *Waste management*, 48, 584–592.

Buhrkal, K., Larsen, A. and Ropke, S. (2012). The waste collection vehicle routing problem with time windows in a city logistics context. *Procedia - Social and Behavioral Sciences*, 39, 241–254.

Calvet, L., de Armas, J., Masip, D. and Juan, A. (2017). Learnheuristics: hybridizing metaheuristics with machine learning for optimization with dynamic inputs. *Open Mathematics*, 15, 261–280.

Calvet, L., Juan, A. A., Serrat, C. and Ries, J. (2016). A statistical learning based approach for parameter fine-tuning of metaheuristics. *SORT-Statistics and Operations Research Transactions*, 1, 201–224.

Clarke, G. and Wright, J. (1964). Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research*, 12, 568–581.

Faulin, J., Juan, A. A., Serrat, C. and Bargueno, V. (2008). Predicting availability functions in time-dependent complex systems with saedes simulation algorithms. *Reliability Engineering & System Safety*, 93, 1761–1771.

Ferone, D., Gruler, A., Festa, P. and Juan, A. A. (2019). Enhancing and extending the classical grasp framework with biased randomisation and simulation. *Journal of the Operational Research Society*, 70, 1362–1375.

Figliozzi, M. A. (2012). The time dependent vehicle routing problem with time windows: Benchmark problems, an efficient solution algorithm, and solution characteristics. *Transportation Research Part E: Logistics and Transportation Review*, 48, 616–636.

Franceschetti, A., Honhon, D., Woensel, T. V., Bektas, T. and Laporte, G. (2013). The time-dependent pollution-routing problem. *Transportation Research Part B: Methodological*, 56, 265–293.

Gendreau, M., Ghiani, G. and Guerriero, E. (2015). Time-dependent routing problems: A review. *Computers & Operations Research*, 64, 189–197.

Ghiani, G., Mourão, C., Pinto, L. and Vigo, D. (2014). Routing in waste collection applications. In: Corberán, A., Laporte, G. (Eds.), *Arc Routing: Problems, Methods, and Applications. SIAM Monographs on Discrete Mathematics and Applications*, Philadelphia, pp. 351–370.

Grasas, A., Juan, A. A., Faulin, J., de Armas, J. and Ramalhinho, H. (2017). Biased randomization of heuristics using skewed probability distributions: a survey and some applications. *Computers & Industrial Engineering*, 110, 216–228.

Gruler, A., Fikar, C., Juan, A. A., Hirsch, P. and Contreras-Bolton, C. (2017a). Supporting multi-depot and stochastic waste collection management in clustered urban areas via simulation–optimization. *Journal of Simulation*, 11, 11–19.

Gruler, A., Quintero-Araujo, C., Calvet, L. and Juan, A. A. (2017b). Waste collection under uncertainty: a simheuristic based on variable neighborhood search. *European Journal of Industrial Engineering*, 11.

Han, H. and Ponce-Cueto, E. (2015). Waste collection vehicle routing problem: literature review. *Traffic & Transportation*, 27, 345–358.

Hannan, M., Akhtar, M., Begum, R. A., Basri, H., Hussain, A. and Scavino, E. (2018). Capacitated vehicle-routing problem model for scheduled solid waste collection and route optimization using pso algorithm. *Waste management*, 71, 31–41.

Harwood, K., Mumford, C. and Eglese, R. (2013). Investigating the use of metaheuristics for solving single vehicle routing problems with time-varying traversal costs. *Journal of the Operational Research Society*, 64, 34–47.

Hashimoto, H., Yagiura, M. and Ibaraki, T. (2008). An iterated local search algorithm for the time-dependent vehicle routing problem with time windows. *Discrete Optimization*, 5, 434–456.

Hemmelmayr, V., Doerner, K. F., Hartl, R. F. and Rath, S. (2013). A heuristic solution method for node routing based solid waste collection problems. *Journal of Heuristics*, 19, 129–156.

Hill, A. V. and Benton, W. C. (1992). Modelling intra-city time-dependent travel speeds for vehicle scheduling problems. *Journal of the Operational Research Society*, 43, 343–351.

Ichoua, S., Gendreau, M. and Potvin, J.-Y. (2003). Vehicle dispatching with time-dependent travel times. *European Journal of Operational Research*, 144, 379–396.

Ismail, Z. and Irhamah, I. (2008). Solving the vehicle routing problem with stochastic demands via hybrid genetic algorithm-tabu search. *Journal of Mathematics and Statistics*, 4, 161–167.

Ismail, Z. and Loh, S. (2009). Ant colony optimization for solving solid waste collection scheduling problems. *Journal of Mathematics and Statistics*, 5, 199–205.

Jabali, O., Van Woensel, T. and de Kok, A. (2012). Analysis of travel times and CO2 emissions in time-dependent vehicle routing. *Production and Operations Management*, 21, 1060–1074.

Juan, A. A., Kelton, W. D., Currie, C. S. and Faulin, J. (2018). Simheuristics applications: dealing with uncertainty in logistics, transportation, and other supply chain areas. In: *2018 Winter Simulation Conference (WSC). IEEE*, pp. 3048–3059.

Kim, B., Kim, S. and Sahoo, S. (2006). Waste collection vehicle routing problem with time windows. *Computers & Operations Research*, 33, 3624–3642.

Kok, A. L., Hans, E. W. and Schutten, J. M. J. (2012). Vehicle routing under time-dependent travel times: The impact of congestion avoidance. *Computers & Operations Research*, 39, 910–918.

Kuo, Y. (2010). Using simulated annealing to minimize fuel consumption for the time-dependent vehicle routing problem. *Computers and Industrial Engineering*, 59, 157–165.

Lecluyse, C., Van Woensel, T. and Peremans, H. (2009). Vehicle routing with stochastic time-dependent travel times. *4OR*, 7, 363.

Lourenço, H. R., Martin, O. C. and Stützle, T. (2003). Iterated local search. In: Glover, F., Kochenberger, G. A. (Eds.), *Handbook of Metaheuristics*. Springer US, Boston, MA, pp. 320–353.

Malandraki, C. and Daskin, M. S. (1992). Time dependent vehicle routing problems: Formulations, properties and heuristic algorithms. *Transportation Science*, 26, 185–200.

Markov, I., Varone, S. and Bierlaire, M. (2016). Integrating a heterogeneous fixed fleet and a flexible assignment of destination depots in the waste collection VRP with intermediate facilities. *Transportation Research Part B: Methodological*, 84, 256–273.

Mohsenizadeh, M., Tural, M. K. and Kentel, E. (2020). Municipal solid waste management with cost minimization and emission control objectives: A case study of ankara. *Sustainable Cities and Society*, 52, 101807.

Muyldermans, L., Beullens, P., Cattrysse, D. and Van Oudheusden, D. (2005). Exploring variants of 2-opt and 3-opt for the general routing problem. *Operations Research*, 53, 982–995.

Nahum, O. E. and Hadas, Y. (2009). Developing a model for the stochastic time-dependent vehicle-routing problem. In: *2009 International Conference on Computers Industrial Engineering*. pp. 118–123.

Nolz, P., Absi, N. and Feillet, D. (2014). A stochastic inventory routing problem for infectious medical waste collection. *Networks*, 63, 82–95.

Nuortio, T., Kytöjoki, J., Niska, H. and Bräysy, O. (2006). Improved route planning and scheduling of waste collection and transport. *Expert Systems with Applications*, 30, 223–232.

Pillac, V., Gendreau, M., GuÃ©ret, C. and Medaglia, A. L. (2013). A review of dynamic vehicle routing problems. *European Journal of Operational Research*, 225, 1–11.

Quintero-Araujo, C. L., Caballero-Villalobos, J. P., Juan, A. A. and Montoya-Torres, J. R. (2017). A biased-randomized metaheuristic for the capacitated location routing problem. *International Transactions in Operational Research*, 24, 1079–1098.

Ramos, T. R. P., Gomes, M. I. and Barbosa-Póvoa, A. P. (2014). Economic and environmental concerns in planning recyclable waste collection systems. *Transportation Research Part E: Logistics and Transportation Review*, 62, 34–54.

Resende, M. G. and Ribeiro, C. C. (2010). Greedy randomized adaptive search procedures: Advances, hybridizations, and applications. In: *Handbook of metaheuristics*. Springer, pp. 283–319.

Ritzinger, U., Puchinger, J. and Hartl, R. F. (2016). A survey on dynamic and stochastic vehicle routing problems. *International Journal of Production Research*, 54, 215–231.

Soysal, M., Bloemhof-Ruwaard, J. M. and Bektas, T. (2015). The time-dependent two-echelon capacitated vehicle routing problem with environmental considerations. *International Journal of Production Economics*, 164, 366–378.

Strand, M., Syberfeldt, A. and Geertsen, A. (2020). A decision support system for sustainable waste collection. In: *Waste Management: Concepts, Methodologies, Tools, and Applications. IGI Global*, pp. 347–365.

Tadei, R., Perboli, G. and Perfetti, F. (2017). The multi-path traveling salesman problem with stochastic travel costs. *EURO Journal on Transportation and Logistics*, 6, 3–23.

Tas, D., Dellaert, N., van Woensel, T. and de Kok, T. (2014). The time-dependent vehicle routing problem with soft time windows and stochastic travel times. *Transportation Research Part C: Emerging Technologies*, 48, 66–83.

Teixeira, J., Antunes, A. P. and de Sousa, J. P. (2004). Recyclable waste collection planning-a case study. *European Journal of Operational Research*, 158, 543–554.

Zhang, T., Chaovalitwongse, W. A. and Zhang, Y. (2014). Integrated ant colony and tabu search approach for time dependent vehicle routing problems with simultaneous pickup and delivery. *Journal of Combinatorial Optimization*, 28, 288–309.

SORT 44 (2) July-December 2020, 311-334 DOI: 10.2436/20.8080.02.104

Why simheuristics? Benefits, limitations, and best practices when combining metaheuristics with simulation

Manuel Chica[1,2], Angel A. Juan[3], Christopher Bayliss[3], Oscar Cordón[1] and W. David Kelton[4,5]

Abstract

Many decision-making processes in our society involve NP-hard optimization problems. The large-scale, dynamism, and uncertainty of these problems constrain the potential use of stand-alone optimization methods. The same applies for isolated simulation models, which do not have the potential to find optimal solutions in a combinatorial environment. This paper discusses the utilization of modelling and solving approaches based on the integration of simulation with metaheuristics. These 'simheuristic' algorithms, which constitute a natural extension of both metaheuristics and simulation techniques, should be used as a 'first-resort' method when addressing large-scale and NP-hard optimization problems under uncertainty –which is a frequent case in real-life applications. We outline the benefits and limitations of simheuristic algorithms, provide numerical experiments that validate our arguments, review some recent publications, and outline the best practices to consider during their design and implementation stages.

MSC: 68U20, 68T20, 90C59, 90C27.

Keywords: Simulation, metaheuristics, combinatorial optimization, simheuristics.

1 Introduction

Decision makers in areas such as transportation, logistics, supply-chain management, health care, production, telecommunication systems, and finance have to face complex challenges when tackling optimization problems in real-world applications. Most of these optimization problems are NP-hard, while others have a lack of complete information that makes their exact definition or formulation quite challenging if not impossible. These facts limit the use of exact optimization methods to small- and medium-sized

[1] Andalusian Research Institute DaSCI, University of Granada, Granada, Spain.

[2] School of Electrical Eng. and Computing, The University of Newcastle, Callaghan NSW, Australia.

[3] IN3 – Computer Science Department, Universitat Oberta de Catalunya, Barcelona, Spain.

[4] Department of Operations, BA, and Inf. Systems, Univ. of Cincinnati, Cincinnati, OH, USA.

[5] Department of Operations Research, Naval Postgraduate School, Monterey, CA, USA.

Received: September 2018
Accepted: July 2020

instances, in which the optimal values can be obtained in reasonable computing times. Moreover, traditional optimization methods might require the use of simplifying assumptions, which do not always reflect the actual system characteristics in a proper manner. Driven by economic and technological factors, real-world systems are becoming increasingly large and complex. Among these factors, we could include trends such as globalization, increased computing power, information technologies, as well as the availability of vast amounts of data (Xu et al., 2015).

Metaheuristic algorithms have gained popularity as a predominant approach for solving real-world optimization problems (Dokeroglu et al., 2019). These algorithms are able to deal with non-trivial objective functions (e.g., multi-objective, non-convex, non-smooth, and noisy functions), soft constraints, and decision variables of different nature. Metaheuristics allow decision makers to obtain near-optimal solutions to large and complex problems in reasonably low computing times, sometimes even in real time (e.g., a few seconds). Therefore, they have become effective methodologies in application areas where optimization of system resources is needed. In addition, approaches hybridizing exact methods with metaheuristics are also widely used. For instance, matheuristics (Boschetti et al., 2009) combine both approaches to get the best from each of them. Typically, they employ the metaheuristic component to deal with the large global problem, while the exact component is used to cope with specific parts of it (Fischetti and Fischetti, 2018). Nonetheless, both exact optimization methods and metaheuristics frequently assume that the problem inputs, the underlying objective functions, and the set of optimization constraints are deterministic or follow simple probabilistic rules. These are strong assumptions and, as a consequence, many deterministic models are oversimplified versions of real-world systems. Coping with the inherent uncertainty of the systems to optimize during problem solving has recently gained relevance (Keith and Ahner, 2019). For instance, robust approaches for metaheuristics have been proposed to handle such uncertainty (Beyer and Sendhoff, 2007). Most of these approaches are extensions of exact optimization models, and they can be classified as deterministic (i.e., based on a set of plausible scenarios), probabilistic (i.e., assuming a given probabilistic function), or possibilistic (i.e., fuzzy-interval measures).

Simulation can be understood as the process of model 'execution' that takes a model through its evolution over time. This evolution can produce changes in the system state or not (stationary system). In addition, these changes can occur discretely or continuously through time. In discrete simulation, the event-oriented view works with the logic occurring at the instantaneous discrete events themselves, rather than with entities and resources (Wainer, 2017). However, the process-oriented world-view describes how entities move through various processes, where each process may require one or more resources and takes a certain (usually stochastic) amount of time (Couture et al., 2018). Simulation allows us to represent the real system in detail and can maintain better control over experimental conditions than by experimenting with the real system itself. A simulation model can be defined as a set of rules (e.g., equations, flowcharts, or state machines) that define how the system evolve in the future and how uncertain the system

is at its present state. A valid simulation model might be able to capture the existing complex reality in a realistic and precise way. A well validated simulation should be one of the preferred approaches to employ when modelling uncertainty in real-world complex optimization problems. As Lucas et al. (2015) noted, "simulation is now an option that should be, in many ways, regarded as the method of choice for analysing complex systems in the face of astounding advances in affordable processing power, modelling paradigms and tools, and supporting analysis capabilities". Still, stand-alone simulation methods show limitations when dealing with optimization problems of combinatorial nature, since a classical simulation approach does not incorporate efficient search methods to explore vast solution spaces.

Hence, both simulation-optimization (Fu, 2015) and simulation-based optimization (Gosavi, 2015) methods can provide practitioners with a flexible and rich tool when dealing with optimization problems in uncertain domains. In particular, we focus here on a subset of these methods that uses metaheuristics for the optimization component. When properly designed, these 'simheuristics' are capable of solving NP-hard and stochastic optimization problems where the simulation component copes with the uncertainty of the system and interacts with the metaheuristic component (Juan et al., 2018). The latter component, in turn, searches the solution space for a near-optimal result. In the past, some optimization problems have been solved by using simulation to evaluate the quality of solutions in engineering. Notice, however, that simheuristic algorithms go one step beyond in the sense that: (i) the feedback from the simulation should also be used to guide the metaheuristic search process itself; and (ii) all the information provided by the simulation component for a solution to the stochastic optimization problem (stochastic solution) allows considering a risk / reliability analysis; then, this analysis can be used to assess alternative stochastic solutions to the stochastic optimization problem. All these characteristics, plus the fact that integration of simulation techniques with metaheuristic algorithms is relatively simple, make simheuristics a 'first-resort' method when dealing with real-world optimization problems under uncertainty conditions. In this paper, we analyse some of the advantages of using simheuristics over traditional methods, as well as some of their limitations. Advantages range from a better understanding of the system behaviour to the use of the generated information through the different simheuristic stages. For example, visualization, machine learning, and sensitivity analysis can be easily used to obtain richer information about the optimization process. We also describe how this combination of metaheuristics and simulation can be carried out to build a successful simheuristic. Several construction guidelines are given to help researchers and practitioners reach their goals. Thus, for instance, validation and stakeholders' discussion of the simulation model used within the simheuristic design and testing stages are encouraged. As simulation can tolerate far less restrictive modelling assumptions, even simple simulations must be correctly validated (Chica et al., 2017) and agreed to by as many decision makers as possible in order to lead to better decisions (Voinov and Bousquet, 2010). These guidelines promote the use of different stages to avoid jeopardizing the optimization process itself, thus obtain the best possible results with reduced com-

puting times. The paper also includes some computational experiments that contribute to support our claims, as well as a number of references to recent publications with additional numerical results. These 'auxiliary' references show applications of simheuristics to different fields.

The rest of the paper is structured as follows: Section 2 provides a short overview of metaheuristic algorithms. Section 3 discusses how uncertainty has been traditionally addressed in optimization problems. Section 4 analyses the basic concepts behind a simheuristic approach. Section 5 reviews previous simheuristic applications in terms of their constituent components and general results. Section 6 lists the most important advantages of using simheuristics, while Section 7 studies their main limitations and how they can be partially overcome. Section 8 provides some guidelines that can be useful during the design and implementation stages of a simheuristic algorithm. Finally, concluding remarks are provided in Section 9.

2 An overview on metaheuristic optimization

According to Glover and Kochenberger (2006), metaheuristics can be defined as "an iterative process that guides the operation of one or more subordinate heuristics (which may be from a local search process to a constructive process of random solutions) to efficiently produce quality solutions for a problem". Metaheuristics are a family of approximate non-linear optimization techniques that provide acceptable solutions (typically near-optimal ones), in a reasonable amount of time, for solving computationally hard and complex problems in science, engineering, and other fields. Unlike exact optimization algorithms, metaheuristics do not guarantee provably optimal solutions. However, for many large-scale real-world problems, metaheuristics might be preferred over gradient-based methods or mathematical programming (Singh and Jana, 2017). The same is true in the case of optimization problems with non-smooth objective functions (Juan et al., 2020). There are also effective gradient-based methods, like the simultaneous perturbation stochastic approximation one (Spall, 2005). These methods are suitable for adaptive modelling and optimization under uncertainty (Bhatnagar et al., 2003) and control optimization (Li, Jafarpour and Mohammad-Khaninezhad, 2013). However, these methods show limitations in the presence of non-smooth objective functions (like the ones due to the existence of realistic soft constraints), where gradients cannot be easily computed. Metaheuristics, on the other hand, are derivative-free optimization methods.

Metaheuristics can be classified according to various characteristics (Talbi, 2009): nature-inspired vs. not nature-inspired, deterministic vs. stochastic, population-based vs. single-solution, iterative vs. greedy, etc. Another issue to be taken into account when selecting a metaheuristic is its exploration versus exploitation capabilities. This concept is usually linked to different sub-families. Thus, while single-solution-based algorithms manipulate and transform a single solution during the search (high intensification),

population-based algorithms evolve a whole population of solutions (high diversification). Single-solution-based metaheuristics could be viewed as 'walks' through neighbourhoods or search trajectories across the search space of the problem at hand. They are performed by iterative procedures that move from the current solution to another one based on local search methods. Among others, some of the most prominent metaheuristics of this sub-family are: tabu search (Glover and Laguna, 2013), simulated annealing (Kirkpatrick, Gelatt and Vecchi, 1983), variable neighbourhood search (Hansen, Mladenovic and Moreno, 2010), the greedy randomized adaptive search procedure, or GRASP (Feo and Resende, 1995), and iterated local search (Lourenço, Martin and Stutzle, 2010). Within the set of population-based metaheuristics, evolutionary algorithms and, in particular, genetic algorithms are frequently used in many engineering and production problems (Lee, 2018). There are many other algorithms that are based on handling a set of solutions at every iteration. These are ant-colony optimization (Dorigo and Stützle, 2004), particle-swarm optimization (Kennedy, 2010), scatter search (Laguna and Marti, 2012), and estimation of distribution algorithms (Larranaga and Lozano, 2002), among others. Finally, memetic algorithms (Moscato and Mathieson, 2019) can be seen as a marriage between population-based metaheuristics and single-solution metaheuristics. A recent and complete review on metaheuristics can be found in Hussain et al. (2019).

3 Handling with uncertainty in optimization problems

The traditional formulation of optimization problems is inherently static and deterministic. However, reality is dynamic and uncertain: environmental parameters fluctuate, materials wear down, processing or transportation times vary, clients change their demands, etc. (Beyer and Sendhoff, 2007). When uncertainty is absent from the optimization formulation, the optimized solutions for those systems may be unstable and sensitive to small changes in the input parameters. A traditional way to tackle this uncertainty in optimization is by providing a high degree of robustness in the solutions. In optimization problems, robust solutions are those that remain relatively unchanged when exposed to uncertainty. Thus, a robust solution can be seen as one which is less sensitive to the perturbation of their environmental or operating conditions, uncertainties in the model outputs, and / or imprecision when measuring the decision variables. Strictly speaking, robust solutions are guaranteed to remain insensitive to changes in the system –at least within a certain range. Recoverable robustness requires that a solution is recoverable in all outcomes. Beyond these definitions, there are more relaxed and attainable degrees of robustness. In general, a robust solution possesses some specified minimum level of reliability or performance level over all outcomes and eventualities (Faulin et al., 2008). Taguchi (1989) envisioned a three-stage design methodology for robust optimization: the system, parameters, and tolerance designs. In Taguchi's method, there are two main classes of optimization parameters: (i) controllable parameters x that are to be tuned;

and (ii) uncontrollable noise factors ξ, such as environmental conditions or production tolerances. In a real-world system, an optimal design has to face different types of robustness depending on the source of uncertainties on the latter parameters: changing environmental and operating conditions, production tolerances and actuator imprecision, uncertainties in the system output, and feasibility uncertainty. These types of uncertainties are usually handled by optimization methods in three different ways (Beyer and Sendhoff, 2007): deterministic, probabilistic, and possibilistic. A common approach followed in robust optimization is to consider the worst-case scenario. However, this is a conservative approach since it can result in poor optimization performance, and even in a solution that is useless in reality. Another methodology is to consider a predefined set of deterministic scenarios, where some of the parameters of the problem are uncertain or depend upon future actions (Chica et al., 2016). As an extension of this approach, an associated probability distribution could be assigned to each of these potential scenarios. Also, the search for optimal robust designs often appears as a multi-criteria decision problem, e.g.: while optimizing a conditional expectation and a large dispersion or variance. In all these cases there is a trade-off between maximal expected performance and variance. For example, one proposal along these lines is the multi-objective six sigma of Shimoyama, Oyama and Fujii (2005), who define robustness as "stability of the system against uncertainty".

Simulation-optimization methods in general (Fu, 2002), and simulation-based optimization in particular (Gosavi, 2015) constitute an excellent choice to deal with optimization problems with stochastic components. Modern computing hardware, modelling paradigms, and advanced simulation software have together made these approaches the methods of choice that can produce results to complex stochastic problems, which cannot be easily and efficiently addressed using more traditional methodologies. Simulation optimization has benefited from the development of both general computing, metaheuristics, stochastic programming, and simulation-specific modelling paradigms. Thus, simulation-optimization methods –which include simulation-based optimization and simheuristics, among others– might be an excellent choice when solving complex problems where time dynamics and uncertainty are important. Simheuristics (Juan et al., 2018) can be seen as a particular type of simulation-based optimization. Combining metaheuristics with simulation models is becoming popular as an effective procedure to deal with complex combinatorial optimization problems. To the best of our knowledge, it was with the work of Glover, Kelly and Laguna (1996, 1999) and April et al. (2003) where this combination was popularized. These authors were the promoters of OptQuest, a 'black-box' optimum-seeking software product that is currently integrated into several commercial simulation-modelling packages. By using this commercial software in concert with simulation-modelling packages, a stochastic simulation model is developed for a given system. Then, the input parameters of interest are changed in an attempt to optimize a designated output performance metric (Kleijnen and Wan, 2007).

To end this section, one should mention other approaches that are also used to deal with stochastic optimization problems. One of the most popular is stochastic program-

ming (Prékopa, 2013). Stochastic programming integrates uncertainty consideration in mathematical programming models. This approach might be highly efficient when considering multi-stage decision processes with a reduced number of possible scenarios at each stage. However, it might also have scalability issues as the number of scenarios and stages grows. The literature on stochastic programming is quite huge, so the interested reader is referred to Ruszczyński and Shapiro (2003) for a nice overview of stochastic programming models. Similarly, stochastic Petri nets (Tigane, Kahloul and Bourekkache, 2017) provide a powerful set of building blocks for specifying the state-transition mechanism and event-scheduling mechanism of a discrete-event stochastic system. These nets are well suited to represent concurrency, synchronization, precedence, and priority phenomena. As such, they have been used in optimization problems under uncertainty scenarios (Melani et al., 2019). Finally, chaos theory allows analysing patterns of outcomes over time that evolve according to a deterministic equation, with these outcomes being extremely sensitive to the initial conditions. This paradigm allows for the modelling of events that are unexpected, i.e.: 'black swan' events (Taleb and Swan, 2008). Chaos theory can be combined with optimization techniques to address stochastic optimization problems (Anter and Ali, 2020).

4 The simheuristic approach

As discussed in Hubscher-Younger et al. (2012), it is not always possible to apply a simulation-optimization software directly out of the box. Instead, it needs to be adapted to the specific characteristics of the problem. Thus, researchers in the optimization community proposed more flexible and 'white-box' approaches. Basically, simheuristics make use of a simulation paradigm to extend existing and efficient metaheuristics. As metaheuristics are primarily designed to cope with deterministic problems, simheuristics can be seen as a metaheuristic extension to be employed when solving optimization problems under uncertainty. This simheuristics approach can be considered a subset of the simulation-for-optimization paradigm. For example, Andradóttir (2006) elaborates on the subject of simulation-based optimization methods, providing a survey on optimization add-ons for discrete-event simulation software. As pointed out by Figueira and Almada-Lobo (2014), simulation-optimization methods are designed to combine the best of both approaches in order to deal with: (i) optimization problems with stochastic components; and (ii) simulation models with optimization requirements. Among these simulation-optimization methods, the combination of simulation with metaheuristics is a promising approach for solving stochastic optimization problems that are frequently encountered by decision makers in the aforementioned industrial sectors (Glover et al., 1996, 1999). A discussion on how random search can be incorporated in simulation-optimization approaches is provided in Andradóttir (2006), while reviews and tutorials on simulation-optimization can be found in Chau et al. (2014) and Jian and Henderson (2015). Likewise, simheuristics can be seen as a specialized case of simulation-

based optimization (April et al., 2003). Hybridization of simulation techniques with metaheuristics allows us to consider stochastic variables in the objective function of the optimization problem, as well as probabilistic constraints in its mathematical formulation. Hence, a simheuristic algorithm contains a particular simulation for an optimization approach, and it is oriented efficiently to tackle an optimization problem involving stochastic components. These stochastic components can be either located in the objective function (e.g., random customers' demands, random processing times, etc.) or in the set of constraints (e.g., customers' demands that must be satisfied with a given probability, deadlines that must be met with a given probability, etc.). Therefore, most of the metaheuristic frameworks can be easily extended to simheuristics, as discussed in Ferone et al. (2019) for the GRASP. For this reason, when dealing with large-scale NP-hard optimization problems –where uncertainty is present–, researchers should consider simheuristics as a 'first-resort' method, since they empower metaheuristic approaches to cope with more realistic stochastic models.

While exact and analytical methods offer superior performance in the optimality dimension (i.e., the capacity to reach optimal values), they have severe limitations in other relevant dimensions such as scalability (i.e., ability to deal with large-scale problems), modelling (i.e., capacity to develop models that accurately represent the real-life system), uncertainty (i.e., ability to cope with non-deterministic scenarios), or computing times (especially for large-scale instances of complex optimization problems). Being an offspring of metaheuristics and simulation, simheuristics inherits the best properties of both methodologies, thus extending metaheuristics so they can deal with uncertainty. At the same time, by adding a metaheuristic optimization component, they also extend simulation methods with the capability of coping with optimization problems successfully. Seminal research on these concepts showed applications of this methodology to different fields. Thus, for instance, April et al. (2006) constructed a simheuristic based on a discrete-event simulation model of a hospital emergency room. Their goal was to determine the optimal configuration of resources that results in the shortest average length of stay for patients. These authors also developed a simulation-optimization algorithm to minimize staffing levels for personal claims processing in an insurance company. Juan et al. (2011) employed a basic simheuristic to deal with the vehicle routing problem with stochastic demands. An enhanced and extended version of their approach was developed by Calvet et al. (2019) to solve the multi-depot stochastic vehicle routing problem. Juan et al. (2014) used a simheuristic to solve the single-period inventory routing problem with stochastic demands and stock outs, while Gruler et al. (2020a) extended the previous approach to the stochastic multi-period inventory routing problem. Gonzalez-Neira et al. (2017) and Hatami et al. (2018) presented simheuristic approaches for solving different permutation flow-shop problems with stochastic processing times. An example of simheuristic applications to distributed computer networks can be found in Cabrera et al. (2014), where discrete-event simulation is combined with a simple metaheuristic framework to optimize a very large, dynamic network of non-dedicated computers offering online services over the Internet. Gruler et al. (2017a, 2020b) developed simheuristic

approaches for supporting stochastic waste-collection management in urban areas. In De Armas et al. (2017), the authors extended a metaheuristic approach into a simheuristic one in order to cope with a stochastic version of the facility location problem. Gruler et al. (2019) propose the use of simheuristics to model human network behaviour. Finally, Reyes-Rubiano et al. (2019) introduce a simheuristic algorithm for solving the electric vehicle routing problem with stochastic travel times. Most of the aforementioned applications refer to the integration of Monte Carlo (MC) simulation with a metaheuristic framework. However, other simulation paradigms are also possible (Rabe, Deininger and Juan, 2020). Overall, we distinguish four main simulation paradigms to be used within a simheuristic. Apart from MC simulation, discrete event simulation (Heath et al., 2011), system dynamics (Sterman, 2001), and agent-based modelling (Kasaie and Kelton, 2015) are specially suitable depending on the optimization-problem characteristics and available resources.

5 Analysis of existing work and some numerical results

In this section, we reflect on the previous implementations of simheuristics and begin by analysing the structure of simheuristics when they are applied to different problem domains. We also consider possible future simheuristic developments, and focus on the general results that emerge from simheuristic algorithms applied to different fields. Likewise, similarities that exist among simheuristic applications are also discussed, as well as the evolution of the simheuristic framework. Firstly, each simheuristic has the following common steps: (i) an input deterministic equivalent model of the stochastic combinatorial optimization problem; (ii) an iterative search stage that integrates information from simulation testing of candidate solutions; and (iii) one or several best stochastic solutions (i.e., solutions for the stochastic version of the problem), which are returned as the output at the end of the algorithm. Regarding the variety of cases that arise when considering different problem domains, the following can be said: in some cases, the simulation component – which is typically a Monte Carlo simulation or a discrete-event simulation – is used only in a parameter initialization phase, where the expected costs of some predefined policies are approximated. This is the case, for instance, in which the fixed costs are a function of the decision variables but the stochastic/variable costs are not. In most of the applications considered so far, demand has been the stochastic element. Other applications consider processing time uncertainty, service costs, node availability, and cash flows. The simheuristic framework is easily extensible for multiple stochastic elements. Most simheuristic implementations employ a distinct initial solution procedure. In some cases this is required because the metaheuristic component is, by itself, only capable of considering perturbations of a current base solution. In other applications the initialization procedure is used because it had been found that the quality of the initial solution had a significant impact on the quality of the final solution. In more recent applications, it has become increasingly common to use biased-

randomized greedy constructive algorithms (Quintero-Araujo et al., 2017) to generate initial solutions. One of the advantages of such an approach is that it facilitates the use of multi-start metaheuristics, which guarantee a more comprehensive exploration of the search space in question. A similar trend can be seen in the choice of the metaheuristic algorithm. Early applications tended to consider relatively simple but efficient heuristics. Thus, for example, in Gonzalez-Martin et al. (2018) a randomized savings heuristic for the arc routing problem is utilized. Some simheuristics used a local search algorithm as the metaheuristic component. Others, such as the one in Pagès-Bernaus et al. (2019), used iterated local search. Yet, more recent applications use the more advanced variable neighbourhood search metaheuristic (VNS) framework (Panadero et al., 2020). One of the advantages of VNS algorithms is that they use multiple neighbourhood structures, which improve both the exploration and intensification properties of the search trajectory. Given these considerations, the combination of biased-randomization and a VNS search is a very strong approach for ensuring the quality of the optimization component of a simheuristic. In general, the choice of the specific metaheuristic framework should account for the complexity of the simulation component of the problem, as longer simulation times extend the required run times. In other words, simpler simulation models enable the use of more complex metaheuristic algorithms and vice-versa.

Another recurrent theme in simheuristic algorithms is that of using the deterministic value of a candidate solution as a criteria for determining whether that solution should be tested in the integrated simulation component – i.e., as a potential candidate stochastic solution. In applications where simulation runs are not computationally expensive, all candidate solutions can be tested in the integrated simulation model. In different simheuristic applications, the role of the integrated simulation component varies. In some cases the simulation is used to check whether a candidate solution adheres to a number of arbitrary constraints, such as a minimum reliability level (Cabrera et al., 2014). However, by far the most common purpose of the simulation component is that of estimating the stochastic value of a candidate solution. One of the advantages of the simheuristic framework is that both multiple objectives and arbitrary constraints can be handled easily, so future applications could use more of the information output from simulation runs. On the whole, the simulation component of a simheuristic can be utilized during an initial parameter-estimation stage, an optimization stage, and a reliability-analysis stage. The output of simheuristics takes the form of a best stochastic solution or a pool of elite stochastic solutions. Having a pool of elite solutions can be useful for three reasons: (i) for storing promising stochastic solutions and complete a risk / reliability analysis over them; (ii) for storing a Pareto front of non-dominated solutions –in cases where multiple goals are considered, as in Gruler et al. (2017b); and (iii) for providing decision makers with a range of alternative solutions, so that they might be able to select a solution that satisfies a number of other arbitrary constraints. In general, it can be seen that a simheuristic is built from a number of relatively fixed steps, including the choice of simulation paradigm, metaheuristic methodology, and output type. In addition, simheuristics have seen an increasing number of optional steps, in-

cluding: using simulation to provide initial parameter estimates, the use of a distinct initial solution method, and a final detailed reliability analysis. Recent applications tend to include previously introduced steps whilst introducing new ones.

Having discussed the evolving simheuristic framework in some detail, we now consider their possible future evolution. For instance, the input problem that the metaheuristic component searches directly is always the deterministic equivalent model of the stochastic model, where the stochastic variables are replaced by their means. Another approach that could be tested in future applications is to periodically change the deterministic equivalent model by generating random realizations, according to the respective distributions, of some or all of the stochastic variables. Such an approach provides an additional escape mechanism from local stochastic optima. It could also help to improve the diversity of the final elite solution set. Additionally, this represents an alternative method of integrating simulation within the metaheuristic search process. Another possible extension would be to dynamically adjust the number of simulation runs used in the integrated simulation component. For example, the integrated simulation could be terminated as soon as the confidence interval of its stochastic value falls entirely below that of the current best stochastic solution. Such an approach will benefit the run-time of a simheuristic. Yet another possibility would be to generalize the structure of simheuristic algorithms to the extent that it becomes a decision variable. For example, the structure of a simheuristic could be encoded as an integer string. The first integer could correspond to the choice of the initial solution generation method, the second to the choice of the metaheuristic, and so on. Such an approach adds an additional layer to the search, and would thus be most useful for cases where sufficient time is available for generating a solution. In such an investigation, fair testing can be ensured by setting a simulation budget for each instance of a simheuristic algorithm.

Figure 1 displays the gaps of the best deterministic solutions (those associated with the deterministic version of the problem when they are used in a stochastic environment) and the best stochastic solutions (those associated with the stochastic version) found by different simheuristic algorithms. These gaps are computed with respect to the best-known solution for the deterministic version of the problem when it is assessed in a scenario without uncertainty. From this figure, one can conclude that optimal/near-optimal deterministic solutions might have a poor performance in stochastic scenarios. Notice that this result holds in a wide variety of problem domains. In the following, deterministic scenarios/solutions are denoted as det, while stochastic scenarios/solutions are denoted as $stoch$. For example $OBS_{det,stoch}$ refers to the objective value of our best deterministic solution when evaluated in a stochastic scenario. Then, the Figure also supports the following general result for a minimization problem: $BKS_{det,det} \leq E\left[OBS_{stoch,stoch}\right] \leq E\left[OBS_{det,stoch}\right]$, i.e.: the deterministic value of the best-known deterministic solution ($BKS_{det,det}$) is a lower bound for the stochastic value of the best stochastic solution ($OBS_{stoch,stoch}$). At the same time, the latter has the stochastic value of the best-known deterministic solution ($OBS_{det,stoch}$) as an upper bound. Figure 1 also highlights the potential benefits of employing a simheuristic in problems that feature uncertainty.

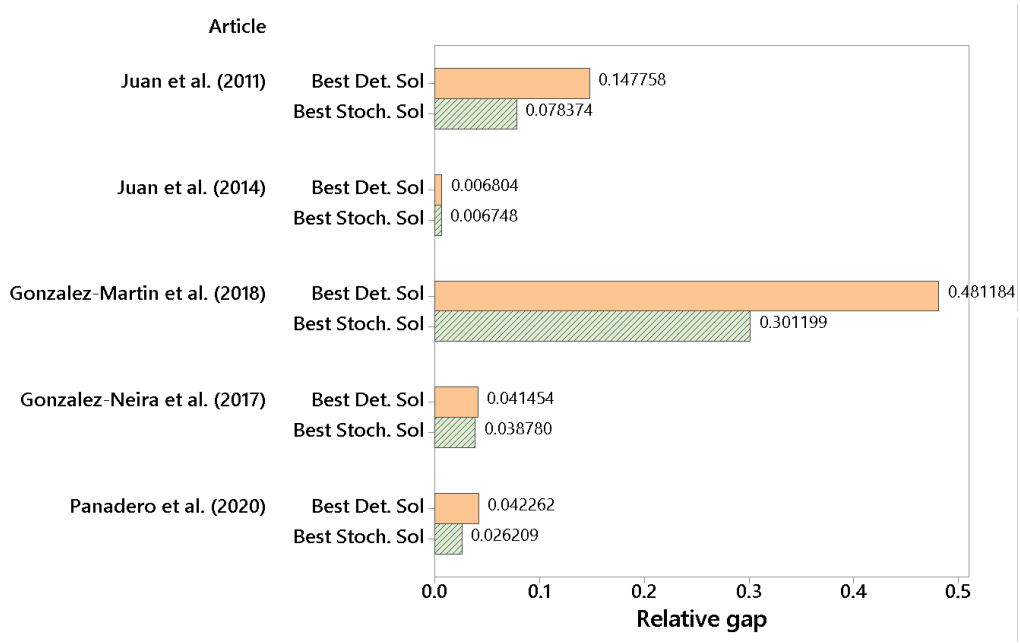


Figure 1: *Relative gaps of the best stochastic and deterministic solutions found by simheuristics compared to the deterministic value of the best-known solutions.*

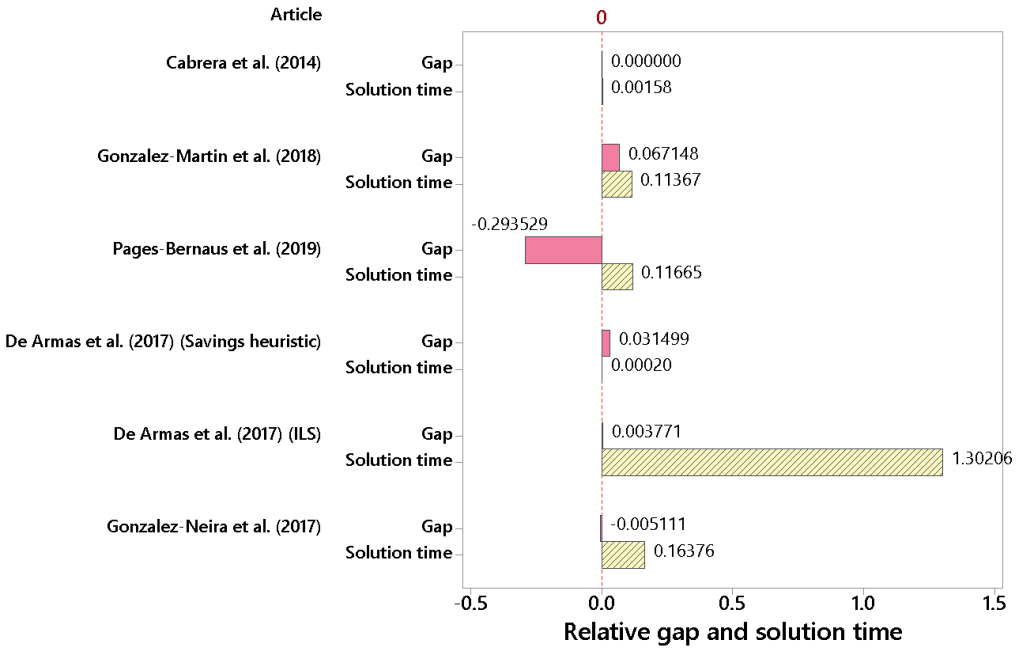


Figure 2: *Optimality gaps and relative solutions times of simheuristics compared to exact formulations over a range of simheuristic applications and problem domains.*

Figure 2 displays optimality gaps and solution times relative to those of several exact methods, for the cases where such experimental results are available. This figure shows that simheuristics are very competitive in terms of the trade-off between solution quality and solution time. Hence, simheuristics are able to generate solutions that are very close to optimality, and can do so in a small fraction of the time required by exact solution approaches.

Likewise, Figure 3 illustrates the effect that the level of variance in the stochastic instance has on the value of the simheuristic solution, as compared with the deterministic value of the best-known solution for the deterministic version of the problem. This figure shows that, in approximately 50% of the cases, increasing the variance of the stochastic parameters of an instance also raises the gap of the stochastic solution relative to the deterministic value of the best-known deterministic solution. In the remaining 50% of the cases, increasing the variance of the stochastic parameters of a problem instance has little or no effect.

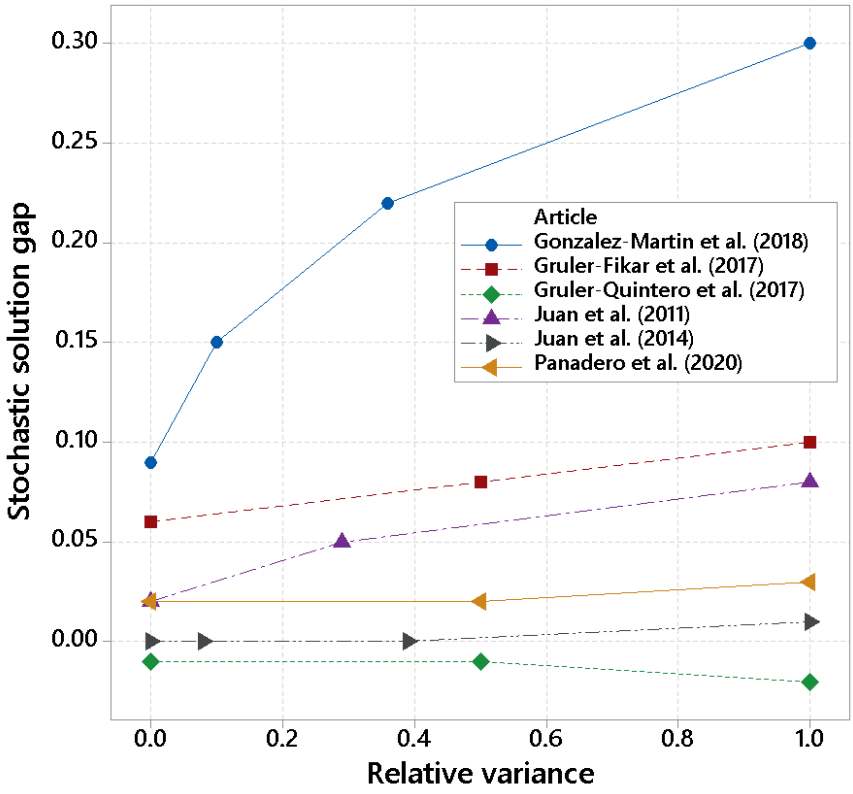


Figure 3: *The effect of increasing the variance of the stochastic variables on the relative gap between the value of our best stochastic solutions and the deterministic value of the best-known deterministic solution.*

6 Advantages of using simheuristics in optimization

This section highlights the main advantages of employing simheuristics, which justify why we propose this methodology as a 'first-resort' method for dealing with optimization problems under uncertainty:

- *Embracing reality by a validated simheuristic:* As opposed to the use of stand-alone analytical models, integrating simulation within a metaheuristic/matheuristic approach allows researchers and decision makers to construct and study valid models of complex systems. Most recent simulation paradigms also allow for analysis of optimization problems under uncertainty with a low number of assumptions. These paradigms also facilitate involvement of stakeholders, who are not directly the modellers of the simheuristic, i.e., participatory modelling (Voinov and Bousquet, 2010). There are new simulation-optimization paradigms that can better represent complex reality, and powerful computational resources to run demanding simulation models. Model validation is a central pillar within the simulation community, as evidenced by its ubiquity in the leading texts over the years (Kelton, Sadowski and Zupick, 2015). But validation should be applied to all modelling, including analytical, so this is not a disadvantage – but a requirement – when using simulation-optimization.

- *Risk assessment of alternative solutions and sensitivity analysis:* Once a simulation is built and validated, finding robust policies and comparing the merits of various policies are two of the main goals (Kleijnen et al., 2005). Joint use of simulation and metaheuristics/matheuristics within a simheuristic framework can help attain these two goals and has advantages compared to other stand-alone methodologies. The results of the simulations can be used to obtain additional information about the probability distribution of the quality of each stochastic solution. This information is then used to introduce a risk/reliability analysis within the decision-making process. The risk-analysis capability of simheuristics is one of its major advantages. This is due to the ability of metaheuritics to generate a set of different solutions, as well as to the ability of the simulation model to provide an observational sampling of the system. Thus, for instance, stochastic solutions with similar expected cost might show different variance, or even different reliability levels; i.e., some routing plans might have a high probability of failure when put into practice, while others might be more reliable. Running a sensitivity analysis (Saltelli et al., 2008) is another advantage of using a simulation together with a metaheuristic method. Sensitivity analysis reveals those input parameters that are most critical in determining the value of key output performance metrics. Usually, this is achieved by exploring the model sensitivity to a particular parameter configuration and input-value options. Sensitivity analysis is typically carried out to gain insights into existing or prospective systems, and this should lead to better decisions and to improved managerial outcomes. This sensitivity analysis can be

directly run by studying the output of the different simulation runs. Although a complete sensitivity analysis requires more advanced methods and specific tools to this end (Chica et al., 2017), the simheuristic learning process can give the modeller a first approach to a deeper sensitivity analysis of the system whose optimization is sought.

- *System understanding and output analysis:* When the simheuristic finishes, we can collect the output-data results and analyse them through machine-learning algorithms to discover hidden properties or relationships. The goal is to enable researchers to identify system patterns interactively, run high-dimensional explorations, or even check the veracity of the approximately-optimized simulation system (Lucas et al., 2015). This is also called the innovization process in evolutionary-computation research (Deb et al., 2014). It means that a set of trade-off optimal or near-optimal solutions, found using metaheuristics, are analysed to decipher useful relationships among problem entities. It provides a better understanding of the problem to a designer or a practitioner. We extend here this concept by adding the simulation face of the simheuristic to enrich the innovization process. Additionally, visualization methods (e.g., histograms, box plots, or scatter plots) can be directly used to visualize post-run simulation outputs that go beyond the traditional analysis of the results. There is an increasing number of studies demonstrating that visualization combined with optimization can promote design innovations and provide decision makers with an improved understanding of the problem (Bonissone, Subbu and Lizzi, 2009). A good visualization enables decision makers to enhance insight into the problem and the different solutions to identify differences and similarities before coming to the final decision (Miettinen, 2014). Exploratory analysis of the input / output variables space of a model is also employed to strengthen confidence in the model realism and to improve understanding of the behaviour of the optimization and simulation models. By analysing the distribution of the model variables and parameters, the modeller can move forward to a simpler and easier-to-understand setting. Use of this exploration, together with sensitivity analysis, provides information on influential factors that significantly affect the variability of the model results, and allow modellers to reach a deeper understanding of the complexity of the model, its uncertainties, interrelationships, and its potential future scenarios (Ligmann-Zielinska et al., 2014).

7 Limitations of Simheuristics

As with any methodology, there are also limitations when using simheuristics. In this section, we highlight some of these limitations as well as some positive aspects that ameliorate their negative impact on the optimum-seeking process.

- *Results are not expected to be truly provably optimal:* Metaheuristics do not ensure an optimal solution to an optimization problem, but rather an acceptable solution in a reasonable amount of time. This fact is amplified when using a simulation to be optimized. Even more, this simulation is a non-linear complex stochastic system that cannot be analytically treated. Therefore, simheuristics are an interesting alternative for practical cases requiring simple and flexible methods that do not need to be globally optimal, although they are usually near-optimal.

- *Additional stakeholders' effort is demanded to define the system*: The set of advantages and 'white-box' paradigms used in a simheuristic also requires additional effort when defining the simulation system and analysing the results provided by the simheuristic. However, we think this design and validation effort is justified as modellers and decision makers can better understand their system from the results of the simheuristics and can adopt the final optimum-seeking results with higher confidence.

- *More computational resources are required compared to traditional methods*: The integration of a simulation engine within a metaheuristic requires high computational effort and also depends on the selected type of simulation paradigm. As will be discussed in Section 8, different strategies can be applied in order to alleviate this effort, such as: (i) 'filtering' the solutions generated by the metaheuristic engine, so that only the 'promising' ones are actually sent to the simulation component; and (ii) using a small number of simulation runs in a first stage, and then analysing in more detail only those that can be classified as 'very promising' solutions.

8 Best design and implementation practices

In this section we outline a set of guidelines or best practices to build a simheuristic algorithm appropriately.

- *Do not overload simheuristics with long simulations:* In general, the modeller has to be careful not to let the simulation jeopardize the computing time given to the entire simulation-metaheuristic process. Otherwise, the metaheuristic would not have time to converge to a good solution if the dimension of the search space is high. Therefore, we recommend decomposing the simheuristic into various stages. For instance, a three-stage approach could be considered. During the first stage, only fast simulations are included in the simheuristic framework. This can be achieved by running the simulation only a limited number of times to obtain rough estimates, or by running the simulation for only those new solutions of the metaheuristic that can be considered as 'promising' ones (e.g., solutions with good

deterministic performance). During this stage, the simulation component of the simheuristic is used not only as a natural way to model the real system, but it also can provide valuable information to the metaheuristic component (i.e., the search process is simulation-driven). For example, it can be used to filter low quality solutions quickly. In a second stage, the best solutions identified in the previous stage are sent throughout a new simulation process with a larger number of iterations to obtain more precise estimates of the uncertain values of the model. The specific number of iterations might be given by error measures such as confidence intervals of the parameters with high uncertainty. Finally, a third and final stage can be used to complete a risk / reliability analysis on the best solutions selected by the decision maker. Dimensions other than the expected value of the solution need to be considered in a high-uncertainty environment, since a solution with a low expected value could also show more variability than other alternative solutions. For example, in a flow-shop scheduling problem with stochastic processing times there might be several solutions (job permutations) that offer a similar expected makespan; however, some of these solutions might show a higher variability than others, or a lower probability of finishing before a given deadline. Similarly, in a vehicle routing problem with stochastic demands, several solutions might offer similar expected costs, but some of these solutions might also show a higher variability than others. Consequently, the decision maker would need more information to decide which solution to choose based on her / his utility function and aversion to risk, or would even need more advanced optimization methods – such as multi-objective optimization – to have a set of solutions with different trade-offs between expected cost value and robust behaviour in the environment.

- *Choose a simulation paradigm that is understandable to decision makers:* Three main goals must be accomplished when developing and selecting the simulation model (Kleijnen et al., 2005): (i) develop a basic understanding of the simulation model and the system it emulates; (ii) find robust policies and decisions; and (iii) compare the merits of various policies or decisions. As mentioned, there is a wide set of available simulation paradigms and within each variant, many variations and possible designs arise. Our guideline here is to use, as much as possible, a participatory simulation-modelling process to increase and share the knowledge and understanding of the system between all the actors involved in the optimization action (Voinov and Bousquet, 2010). This involvement would also clarify and identify the impacts of solutions to a given problem, usually related to the final decision-making support.

- *Choose an appropriate simulation paradigm for each stage of a simheuristic:* Different simulation paradigms can be used for each of the stages of the simheuristic. Then, a more enriched and computationally-intensive simulation model (e.g., agent-based modelling) can be used for the last stages of the simheuristic and

applied only to a reduced set of the solutions provided by the metaheuristic. In contrast, lighter computational simulation models (e.g., a simple Monte Carlo simulation over the stochastic simulation model) might be required in the first stage of the simheuristic. Each individual modelling paradigm has a rich history and exemplar cases in which the strengths of the respective methodology make it a good choice for a particular modelling situation. There also possibilities for combining each pair of approaches to develop hybrid models where each paradigm exploits its strengths (Heath et al., 2011). For instance, Djanatliev and German (2013) present different multi-paradigm simulation methods.

- *Validate the simulation model before running the simheuristic:* A decisive phase when modelling a real-world system is model validation (Oliva, 2003). In our view, this is also a main guideline when designing the simheuristic, as it applies to the simheuristic itself and specifically to its simulation component. The validation requires testing a set of hypotheses, the significance of their behavioural components (by assuming that the behaviour is a consequence of the system structure), and the historical model fitting. Validation is also measured in terms of degrees of confidence or quality, which is usually difficult to obtain for most non-linear simulation models in use (Forrester, 2007). The validation and testing of any model or decision-support system is a decisive step for ensuring its managerial adoption. Decision makers are all rightly concerned about whether results of each model are correct (Sargent, 2005). However, the validation of non-linear models and their effectiveness for real-world problems is not straightforward. The validation stage can be seen as a learning process where the modeller's understanding is enhanced through her / his interaction with the formal and mental model (Morecroft, 2007). As this process evolves, both the formal and mental perceptions of the modellers change, leading to a successive approximation of the formal model to reality. Additionally, the utility and effectiveness of many non-linear models and their outputs are often judged by stakeholders and decision makers (Voinov and Bousquet, 2010). Therefore, it is highly recommended to perform the validation of the models correctly. A set of validation techniques such as calibration (Sargent, 2005), sensitivity analysis (Saltelli et al., 2008), boundary adequacy, and extreme cases tests (Qudrat-Ullah and Seong, 2010) should be carried out for the corresponding simheuristic component in order to guarantee that the simulation model is a valid representation of the underlying system.

9 Concluding remarks

The motivation of this paper is to advocate that a combination of simulation models and metaheuristics / matheuristics should be considered as a first-resort method when dealing with large-scale NP-hard optimization problems with stochastic components,

which is a quite common case when considering real-world challenges. In effect, many real-life optimization problems in areas such as logistics, transportation, scheduling, etc., are complex, large-scale, and involve uncertainties regarding their constraints, input values, and objective functions. Although there are metaheuristic applications that add probabilistic and robustness capabilities to analytical models, they are extensions to the original deterministic model formulation. As we have discussed, integration of simulation methods with metaheuristics and matheuristics is a natural way to cope with these problems. Although prohibitive and unaffordable in the past, advanced simulation methods are now commonly used in research and practice due to widespread and affordable availability of high-performance computing resources and much-improved software for simulation modelling and analysis. The same is true for metaheuristics and matheuristics. As it has been shown in a number of recent publications containing extensive computational experiments, the simheuristics methodology can better face complex reality when seeking optima in uncertain environments.

In this paper we highlighted three main advantages of using simheuristics. First, it is a better way to embrace the reality of the systems we are seeking to optimize. There is no need to include many strong and over-simplifying assumptions to render a tractable model. Second, a simheuristic can easily provide a risk assessment of the optimization-problem solutions. Third, simheuristics facilitate the understanding of the system's behaviour. *A posteriori* analysis applied to the output provided by the simheuristic can help modellers to understand the system dynamics. For instance, one can observe the most sensitive parameters, or even apply statistical analysis to the returned set of optimization solutions to find relationships between them. Visualization techniques are also useful to generate insights about the system, based on the output of the simheuristic method.

Additionally, we have presented the main simulation paradigms to be used within a simheuristic, and a list of guidelines to take into account when designing a simheuristic. We suggested the use of a multi-stage approach to alleviate the required computation effort of the simulation, and the utilization of different simulation paradigms within the simheuristic. Likewise, the need for using a validated simulation model was affirmed. Finally, we encourage the use of a simheuristic paradigm that can be aligned with the 'white-box' paradigm: being understandable and enhancing the decision makers' participation.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science, Innovation, and Universities, and European Regional Development Funds (ERDF) (PID2019-111100RB-C21, RED2018-102642-T, EXASOCO PGC2018-101216-B-I00, and AIMAR (A-TIC-284-UGR18). Likewise, we want to acknowledge the support received by the Erasmus+ programme (2019-I-ES01-KA103-062602). M. Chica is supported by the Ramón y Cajal program (RYC-2016-19800) and SIMARK (PY18-4475), granted by the Andalusian Government.

References

Andradóttir, S. (2006). An overview of simulation optimization via random search. *Handbooks in Operations Research and Management Science*, 13, 617–631.

Anter, A. M. and Ali, M. (2020). Feature selection strategy based on hybrid crow search optimization algorithm integrated with chaos theory and fuzzy c-means algorithm for medical diagnosis problems. *Soft Computing*, 24, 1565–1584.

April, J., Glover, F., Kelly, J. P. and Laguna, M. (2003). Simulation-based optimization: practical introduction to simulation optimization. In *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 71–78. IEEE.

April, J., Better, M., Glover, F., Kelly, J. and Laguna, M. (2006). Enhancing business process management with simulation optimization. In *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 642–649. IEEE.

Beyer, H.-G. and Sendhoff, B. (2007). Robust optimization–a comprehensive survey. *Computer Methods in Applied Mechanics and Engineering*, 196, 3190–3218.

Bhatnagar, S., Fu, M. C., Marcus, S. I. and Wang, I.-J. (2003). Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences. *ACM Transactions on Modeling and Computer Simulation*, 13, 180–209.

Bonissone, P. P., Subbu, R. and Lizzi, J. (2009). Multicriteria decision making: a framework for research and applications. *IEEE Computational Intelligence Magazine*, 4, 48–61.

Boschetti, M. A., Maniezzo, V., Roffilli, M. and Röhler, A. B. (2009). Matheuristics: optimization, simulation and control. In *Hybrid Metaheuristics*, pp. 171–177. Springer.

Cabrera, G., Juan, A. A., Lázaro, D., Marquès, J. M. and Proskurnia, I. (2014). A simulation-optimization approach to deploy internet services in large-scale systems with user-provided resources. *Simulation*, 90, 644–659.

Calvet, L., Wang, D., Juan, A. A. and Bové, L. (2019). Solving the multidepot vehicle routing problem with limited depot capacity and stochastic demands. *International Transactions in Operational Research*, 26, 458–484.

Chau, M., Fu, M. C., Qu, H. and Ryzhov ,I. O. (2014). Simulation optimization: a tutorial overview and recent developments in gradient-based methods. In *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 21–35. IEEE.

Chica, M., Barranquero, J., Kajdanowicz, T., Cordón, O. and Damas, S. (2017). Multimodal optimization: an effective framework for model calibration. *Information Sciences*, 375, 79–97.

Chica, M., Bautista, J., Cordón, Ó. and Damas, S. (2016). A multiobjective model and evolutionary algorithms for robust time and space assembly line balancing under uncertain demand. *Omega*, 58, 55–68.

Couture, R.-M., Moe, S. J., Lin, Y., Kaste, Ø., Haande, S. and Solheim, A. L. (2018). Simulating water quality and ecological status of Lake Vansjø, Norway, under land-use and climate change by linking process-oriented models with a Bayesian network. *Science of the Total Environment*, 621, 713–724.

De Armas, J., Juan, A. A., Marquès, J. M. and Pedroso, J. P. (2017). Solving the deterministic and stochastic uncapacitated facility location problem: from a heuristic to a simheuristic. *Journal of the Operational Research Society*, 68, 1161–1176.

Deb, K., Bandaru, S., Greiner, D., Gaspar-Cunha, A. and Tutum, C. C. (2014). An integrated approach to automated innovization for discovering useful design principles: case studies from engineering. *Applied Soft Computing*, 15, 42–56.

Djanatliev, A. and German, R. (2013). Prospective healthcare decision-making by combined system dynamics, discrete-event and agent-based simulation. In *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 270–281. IEEE.

Dokeroglu, T., Sevinc, E., Kucukyilmaz, T. and Cosar, A. (2019). A survey on new generation metaheuristic algorithms. *Computers & Industrial Engineering*, 137, 106040.

Dorigo, M. and Stützle, T. (2004). *Ant Colony Optimization*. MIT Press, Cambridge.

Faulin, J., Juan, A. A., Serrat, C. and Bargueno, V. (2008). Predicting availability functions in time-dependent complex systems with saedes simulation algorithms. *Reliability Engineering & System Safety*, 93, 1761–1771.

Feo, T. A. and Resende, M. G. C. (1995). Greedy randomized adaptive search procedures. *Journal of Global Optimization*, 6, 109–133.

Ferone, D., Gruler, A., Festa, P. and Juan, A. A. (2019). Enhancing and extending the classical GRASP framework with biased randomisation and simulation. *Journal of the Operational Research Society*, 70, 1362–1375.

Figueira, G. and Almada-Lobo, B. (2014). Hybrid simulation–optimization methods: a taxonomy and discussion. *Simulation Modelling Practice and Theory*, 46, 118–134.

Fischetti, M. and Fischetti, M. (2018). Matheuristics. In *Handbook of Heuristics*, pp. 121–153. Springer.

Forrester, J. W. (2007). System dynamics: the next fifty years. *System Dynamics Review*, 23, 359–370.

Fu, M. C. (2002). Optimization for simulation: theory vs. practice. *INFORMS Journal on Computing*, 14, 192–215.

Fu, M. C. (2015). *Handbook of Simulation Optimization*, Volume 216. Springer.

Glover, F., Kelly, J. P. and Laguna, M. (1996). New advances and applications of combining simulation and optimization. In *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 144–152. IEEE.

Glover, F., Kelly, J. P. and Laguna, M. (1999). New advances for wedding optimization and simulation. In *Proceedings of the Winter Simulation Conference*, Volume 1, Piscataway, New Jersey, pp. 255–260. IEEE.

Glover, F. and Laguna, M. (2013). Tabu search. In *Handbook of Combinatorial Optimization*, pp. 3261–3362. Springer.

Glover, F. W. and Kochenberger, G. A. (2006). *Handbook of Metaheuristics*, Volume 57. Springer Science & Business Media.

Gonzalez-Martin, S., Juan, A. A., Riera, D., Elizondo, M. G. and Ramos, J. J. (2018). A simheuristic algorithm for solving the arc routing problem with stochastic demands. *Journal of Simulation*, 12, 53–66.

Gonzalez-Neira, E. M., Ferone, D., Hatami, S. and Juan, A. A. (2017). A biased-randomized simheuristic for the distributed assembly permutation flowshop problem with stochastic processing times. *Simulation Modelling Practice and Theory*, 79, 23–36.

Gosavi, A. (2015). *Simulation-Based Optimization*. Springer US, New York.

Gruler, A., de Armas, J., Juan, A. A. and Goldsman, D. (2019). Modelling human network behaviour using simulation and optimization tools: the need for hybridization. *SORT-Statistics and Operations Research Transactions*, 43, 0193–222.

Gruler, A., Fikar, C., Juan, A. A., Hirsch, P. and Contreras-Bolton, C. (2017a). Supporting multi-depot and stochastic waste collection management in clustered urban areas via simulation–optimization. *Journal of simulation*, 11, 11–19.

Gruler, A., Panadero, J., de Armas, J., Moreno, J. A. and Juan, A. A. (2020a). A variable neighbourhood search simheuristic for the multiperiod inventory routing problem with stochastic demands. *International Transactions in Operational Research*, 27, 314–335.

Gruler, A., A. Perez-Navarro, L. Calvet, and A. A. Juan (2020b). A simheuristic algorithm for time-dependent waste collection management with stochastic travel times. *SORT-Statistics and Operations Research Transactions*, 44, 1–29.

Gruler, A., Quintero, C. L., Calvet, L. and Juan, A. A. (2017b). Waste collection under uncertainty: a simheuristic based on variable neighbourhood search. *European Journal of Industrial Engineering*, 11, 228–255.

Hansen, P., Mladenović, N. and Moreno, J. A. (2010). Variable neighbourhood search: methods and applications. *Annals of Operations Research*, 175, 367–407.

Hatami, S., Calvet, L., Fernández-Viagas, V., Framiñán, J. M. and Juan, A. A. (2018). A simheuristic algorithm to set up starting times in the stochastic parallel flowshop problem. *Simulation Modelling Practice and Theory*, 86, 55–71.

Heath, S. K., Buss, A., Brailsford, S. C. and Macal, C. M. (2011). Cross-paradigm simulation modelling: challenges and successes. In *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 2788–2802. IEEE.

Hubscher-Younger, T., Mosterman, P. J., DeLand, S., Orqueda, O. and Eastman, D. (2012). Integrating discrete-event and time-based models with optimization for resource allocation. In *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 1–15. IEEE.

Hussain, K., Salleh, M. N. M., Cheng, S. and Shi, Y. (2019). Metaheuristic research: a comprehensive survey. *Artificial Intelligence Review*, 52, 2191–2233.

Jian, N. and Henderson, S. G. (2015). An introduction to simulation optimization. In *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 1780–1794. IEEE.

Juan, A., Faulin, J., Grasman, S., Riera, D., Marull, J. and Mendez, C. (2011). Using safety stocks and simulation to solve the vehicle routing problem with stochastic demands. *Transportation Research Part C: Emerging Technologies*, 19, 751–765.

Juan, A. A., Corlu, C. G., Tordecilla, R. D., de la Torre, R. and Ferrer, A. (2020). On the use of biased-randomized algorithms for solving non-smooth optimization problems. *Algorithms*, 13, 8.

Juan, A. A., Grasman, S. E., Caceres-Cruz, J. and Bektaş, T. (2014). A simheuristic algorithm for the single-period stochastic inventory-routing problem with stock-outs. *Simulation Modelling Practice and Theory*, 46, 40–52.

Juan, A. A., Kelton, W. D., Currie, C. S. and Faulin, J. (2018). Simheuristics applications: dealing with uncertainty in logistics, transportation, and other supply chain areas. In *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 3048–3059. IEEE.

Kasaie, P. and Kelton, W. D. (2015). Guidelines for design and analysis in agent-based simulation studies. In *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 183–193. IEEE.

Keith, A. J. and Ahner, D. K. (2019). A survey of decision making and optimization under uncertainty. *Annals of Operations Research*, SI, 1–35.

Kelton, W. D., Sadowski, R. and Zupick, N. B. (2015). *Simulation with Arena (6th Edition)*. McGraw-Hill Education.

Kennedy, J. (2010). Particle swarm optimization. In *Encyclopedia of Machine Learning*, pp. 760–766. Springer.

Kirkpatrick, S.,Gelatt, J. C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220, 671–680.

Kleijnen, J. P., Sanchez, S. M., Lucas, T. W. and Cioppa, T. M. (2005). State-of-the-art review: a user's guide to the brave new world of designing simulation experiments. *INFORMS Journal on Computing*, 17, 263–289.

Kleijnen, J. P. and Wan, J. (2007). Optimization of simulated systems: OptQuest and alternatives. *Simulation Modelling Practice and Theory*, 15, 354–362.

Laguna, M. and Marti, R. (2012). *Scatter Search: Methodology and Implementations in C*, Volume 24. Springer Science & Business Media.

Larranaga, P. and Lozano, J. A. (2002). *Estimation of Distribution Algorithms: a New Tool for Evolutionary Computation*, Volume 2. Springer Science & Business Media.

Lee, C. K. H. (2018). A review of applications of genetic algorithms in operations management. *Engineering Applications of Artificial Intelligence*, 76, 1–12.

Li, L., Jafarpour, B. and Mohammad-Khaninezhad, M. R. (2013). A simultaneous perturbation stochastic approximation algorithm for coupled well placement and control optimization under geologic uncertainty. *Computational Geosciences*, 17, 167–188.

Ligmann-Zielinska, A., Kramer, D. B., Cheruvelil, K. S. and Soranno, P. A. (2014). Using uncertainty and sensitivity analyses in socioecological agent-based models to improve their analytical performance and policy relevance. *PloS one*, 9, e109779.

Lourenço, H. R., Martin, O. C. and Stützle, T. (2010). Iterated local search: framework and applications. In *Handbook of Metaheuristics*, pp. 363–397. Springer.

Lucas, T. W., Kelton, W. D., Sánchez, P. J., Sanchez, S. M. and Anderson, B. L. (2015). Changing the paradigm: simulation, now a method of first resort. *Naval Research Logistics*, 62, 293–303.

Melani, A. H., Murad, C. A., Caminada Netto, A., Souza, G. F. and Nabeta, S. I. (2019). Maintenance strategy optimization of a coal-fired power plant cooling tower through generalized stochastic Petri nets. *Energies*, 12, 1951.

Miettinen, K. (2014). Survey of methods to visualize alternatives in multiple criteria decision making problems. *OR Spectrum*, 36, 3–37.

Morecroft, J. (2007). *Strategic Modelling and Business Dynamics: A Feedback Systems Approach*. John Wiley & Sons.

Moscato, P. and Mathieson, L. (2019). Memetic algorithms for business analytics and data science: a brief survey. In *Business and Consumer Analytics: New Ideas*, pp. 545–608. Springer.

Oliva, R. (2003). Model calibration as a testing strategy for system dynamics models. *European Journal of Operational Research*, 151, 552–568.

Pagès-Bernaus, A., Ramalhinho, H., Juan, A. A. and Calvet, L. (2019). Designing e-commerce supply chains: a stochastic facility–location approach. *International Transactions in Operational Research*, 26, 507–528.

Panadero, J., Doering, J., Kizys, R., Juan, A. A. and Fito, A. (2020). A variable neighbourhood search simheuristic for project portfolio selection under uncertainty. *Journal of Heuristics*, 26, 353–375.

Prékopa, A. (2013). *Stochastic Programming*, Volume 324. Springer Science & Business Media.

Qudrat-Ullah, H. and Seong, B. S. (2010). How to do structural validity of a system dynamics type simulation model: the case of an energy policy model. *Energy Policy*, 38, 2216–2224.

Quintero-Araujo, C. L., Caballero-Villalobos, J. P., Juan, A. A. and Montoya-Torres, J. R. (2017). A biased-randomized metaheuristic for the capacitated location routing problem. *International Transactions in Operational Research*, 24, 1079–1098.

Rabe, M., Deininger, M. and Juan, A. A. (2020). Speeding up computational times in simheuristics combining genetic algorithms with discrete-event simulation. *Simulation Modelling Practice and Theory*, 103, 102089.

Reyes-Rubiano, L., Ferone, D., Juan, A. A. and Faulin, J. (2019). A simheuristic for routing electric vehicles with limited driving ranges and stochastic travel times. *SORT-Statistics and Operations Research Transactions*, 1, 3–24.

Ruszczyński, A. and Shapiro, A. (2003). Stochastic programming models. *Handbooks in Operations Research and Management Science*, 10, 1–64.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S. (2008). *Global Sensitivity Analysis: the Primer*. John Wiley & Sons.

Sargent, R. G. (2005). Verification and validation of simulation models. In *Proceedings of the Winter simulation Conference*, Piscataway, New Jersey, pp. 130–143. IEEE.

Shimoyama, K., Oyama, A. and Fujii, K. (2005). A new efficient and useful robust optimization approach - design for multi-objective six sigma. In *IEEE Congress on Evolutionary Computation*, Volume 1, Piscataway, New Jersey, pp. 950–957.

Singh, A. and Jana, N. D. (2017). A survey on metaheuristics for solving large scale optimization problems. *International Journal of Computer Applications*, 170, 1–7.

Spall, J. C. (2005). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, Volume 65. John Wiley & Sons.

Sterman, J. D. (2001). System dynamics modelling: tools for learning in a complex world. *California Management Review*, 43, 8–25.

Taguchi, G. (1989). *Introduction to Quality Engineering*. American Supplier Institute.

Talbi, E.-G. (2009). *Metaheuristics: from Design to Implementation*. John Wiley & Sons.

Taleb, N. N. and Swan, B. (2008). *The Impact of the Highly Improbable*. Penguin Books Limited.

Tigane, S., Kahloul, L. and Bourekkache, S. (2017). Reconfigurable stochastic petri nets: A new formalism for reconfigurable discrete event systems. In *International Conference on Mathematics and Information Technology*, pp. 301–308. IEEE.

Voinov, A. and Bousquet, F. (2010). Modelling with stakeholders. *Environmental Modelling & Software*, 25, 1268–1281.

Wainer, G. A. (2017). *Discrete-Event Modeling and Simulation: a Practitioner's Approach*. CRC press.

Xu, J., Huang, E., Chen, C.-H. and Lee, L. H. (2015). Simulation optimization: A review and exploration in the new era of cloud computing and big data. *Asia-Pacific Journal of Operational Research*, 32, 1550019.

SORT 44 (2) July-December 2020, 335-356 DOI: 10.2436/20.8080.02.105

Modelling multivariate, overdispersed count data with correlated and non-normal heterogeneity effects

Iraj Kazemi[1] and Fatemeh Hassanzadeh[2]

Abstract

Mixed Poisson models are most relevant to the analysis of longitudinal count data in various disciplines. A conventional specification of such models relies on the normality of unobserved heterogeneity effects. In practice, such an assumption may be invalid, and non-normal cases are appealing. In this paper, we propose a modelling strategy by allowing the vector of effects to follow the multivariate skew-normal distribution. It can produce dependence between the correlated longitudinal counts by imposing several structures of mixing priors. In a Bayesian setting, the estimation process proceeds by sampling variants from the posterior distributions. We highlight the usefulness of our approach by conducting a simulation study and analysing two real-life data sets taken from the German Socioeconomic Panel and the US Centers for Disease Control and Prevention. By a comparative study, we indicate that the new approach can produce more reliable results compared to traditional mixed models to fit correlated count data.

MSC: 60E05, 62J12, 62J99, 62H20.

Keywords: Bayesian computation, correlated random effects, hierarchical representation, longitudinal data, multivariate skew-normal distribution, over-dispersion.

1 Introduction

An important class of models for count data, in the presence of over-dispersion, is the mixed Poisson. The class includes several popular mixed-Poisson models in terms of choosing mixing priors for unobserved heterogeneity effects. The normal mixing prior was originally introduced by Bulmer (1974) and developed by many others, such as Guo and Trivedi (2002), Miller (2007), and Montesinos et al. (2017) among others. The mixing strategy generates a marginal distribution of longer-tailed than the routinely used Gamma prior, which creates the negative binomial (NB) model (Gonzales-Barron and Butler, 2011). It is also useful in analysing specific over-dispersed count response vari-

[1] Department of Statistics, University of Isfahan, Iran. Tel.: +98-31-37934596.

[2] Department of Statistics, University of Khansar, Iran. Tel.: +98-31-57995312.

Received: December 2019
Accepted: December 2020

ables (Izsák, 2008; Williams and Ebel, 2012). A familiar list of several mixed Poisson distributions is presented by Karlis and Xekalaki (2005), Nadarajah and Kotz (2006a), and Nadarajah and Kotz (2006b). Further models detailed in Kuba and Panholzer (2016) and Cameron and Trivedi (2013).

Count data analysis may involve dealing with both the occurrence of over-dispersion and the correlation between repeated outcomes. A comprehensive overview of the discrete correlated data analysis is provided by Molenberghs, Verbeke and Demetrio (2007) with a discussion on computational issues and the inclusion of many practical applications. In longitudinal studies, the presence of heterogeneity effects is an indication of correlated responses of each subject over time and possibly a sign of over-dispersion. In this scenario, a regular choice to explain variability is the Poisson-multivariate normal (PMN) model, wherein the distribution of effects is assumed to be multivariate normal (e.g., see Chib and Winkelmann, 2001; El-Basyouny and Sayed, 2009; Wu, Deng and Ramakrishnan, 2018). Then, the problem turns to solving an intractable marginal likelihood and requiring advanced computational techniques, such as the Markov chain Monte Carlo (MCMC) in the Bayesian framework.

The associated literature reveals that the multivariate normal is the most adopted mixing prior distribution to the heterogeneity effects. However, it is unlikely to lead always to the best-fitted model. It was our leading motivation to extend the PMN model by setting the multivariate skew-normal mixing prior (Azzalini, 1985; Sahu, Dey and Branco, 2003) for the conditional mean of the Poisson model. The proposed Poisson multivariate skew-normal (PMSN) regression model includes a vector of skewness parameters. Thus we can directly introduce it through an additional hierarchy level to the PMN model. Also, depending on the specific multivariate skew-normal mixing prior, we can define various types of the PMSN model. The proposed model includes Poisson and the PMN as its special cases. Also, the PMSN model reduces to the Poisson skew-normal (PSN) model when unobserved heterogeneity effects are assumed to be independent by introducing a skew-normal mixing prior distribution to the structure of the mixed Poisson model. Specifically, our findings show that the proposed model with various values of the skewness parameter has different performances. In particular, over-dispersion in counts increases as the value of the skewness parameter increases. Results reveal that the PSN over-dispersion is less (more) than the Poisson normal (PN) over-dispersion provided that the skewness parameter being negative (positive). It illustrates that the PSN regression model may be more flexible than the PN model if a count data set exhibits over-dispersion.

From a Bayesian perspective, the proposed models can appear hierarchically to ease the implementation of the Gibbs sampler technique. Also, we use a stochastic representation for the conditional mean of the Poisson regression. It simplifies Bayesian computations due to having the complete conditional posteriors, involved in the Gibbs sampler, in closed forms of known distributions. The Bayesian analysis of correlated count data by fitting the PMN model (e.g., Rizzato et al., 2016) is a specific case of our proposed model. The model fitting is performed by OpenBugs software version 3.2.3, which is

an excellent platform for Bayesian inference using the Gibbs sampler algorithm (e.g., Lunn et al., 2009).

The article is organized as follows. In Section 2, we introduce the PSN model with independent heterogeneity effects for the analysis of count data. In Section 3, mixed-Poisson models with various multivariate skew-normal mixing priors are illustrated for longitudinal count data. In this section, we also emphasize the identification issue in mixed-Poisson models. In Section 4, we present Bayesian mixed models hierarchically to derive the complete conditional posteriors required to implement the Gibbs sampling approach. In Section 5, we conduct a simulation study to compare proposed models with some competing ones. In Section 6, we fit proposed models for the specific data sets taken from follow up studies on the national medical expenditure survey and the polio data. Section 7 gives some concluding remarks.

2 A new modelling methodology to the count data analysis

Assume that the count response Y_{it}, conditioned on the effect u_{it} for subject $i = 1, \cdots, n$ and at time $t = 1, \cdots, T$, follows a Poisson distribution with mean $\exp(\theta_{it})$, where $\theta_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it}$, $\mathbf{x}_{it}$ is a k-dimensional vector of covariates, and $\boldsymbol{\beta}$ is a k-dimensional vector of coefficients. Moreover, the effects u_{it}, defined on the whole real line, are assumed to follow a common probability distribution function (pdf) $G(u_{it}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a vector of parameters that characterize $G(\cdot)$. The marginal density of Y_{it} is called a mixed Poisson density with the probability mass function (pmf) given by integrating out the effects u_{it}. The normal mixing prior for u_{it} leads to the well-known Poisson normal (PN) model. Here, we extend the methodology by letting the mixing prior be skew-normally distributed with the following specification.

Definition 1 *The random variable u_{it}, for subject $i = 1, \cdots, n$ and at time $t = 1, \cdots, T$, follows the skew-normal distribution, denoted by $u_{it} \overset{iid}{\sim} SN\left(\xi, \sigma^2, \delta\right)$, if the density function of u_{it} is given by*

$$g_{SN}\left(u_{it}|\xi, \sigma^2, \delta\right) = 2\varphi\left(u_{it}|\xi, \sigma^2 + \delta^2\right)\Phi\left(\frac{\delta\left(u_{it} - \xi\right)}{\sigma\sqrt{\sigma^2 + \delta^2}}\right), \tag{1}$$

with location parameter $\xi \in \mathbb{R}$, scale parameter $\sigma^2 \in \mathbb{R}^+$ and skewness parameter $\delta \in \mathbb{R}$, where $\varphi(\cdot)$ denotes the pdf of $N\left(\xi, \sigma^2 + \delta^2\right)$ and $\Phi(\cdot)$ denotes the cumulative density function (cdf) of the standard normal (Azzalini, 1985; Sahu et al., 2003).

Using usual statistical methods the following basic properties of density (1) hold.

Properties 1

i. For $\delta = 0$, the original normal mixing prior is retrieved; for $\delta > 0$, positively skewed and for $\delta < 0$, negatively skewed mixing priors are obtained. Figure *1* confirms these results.
ii. The hierarchical representation of u_{it} is shown to be $u_{it}|z_{it} \overset{ind}{\sim} N\left(\xi + \delta z_{it}, \sigma^2\right)$ with $Z_{it} \overset{iid}{\sim} HN(0,1)$, where HN denotes the half-normal distribution. This property helps us to generate a random variable that follows the skew-normal distribution and consequently to implement the McMC approach easily.
iii. The r-th moment of $w_{it} = \exp(u_{it})$, for any real r, is finite and equivalent to the moment generating function (MGF) of the skew-normal distribution. This is explicitly given by $m_r = E(w_{it}^r) = 2\Phi(\delta r)\, exp(r\xi + \frac{1}{2}r^2\left(\sigma^2 + \delta^2\right))$. In particular, the mean and variance of w are $\mu_w = m_1$ and $\sigma_w^2 = m_2 - m_1^2$, respectively.

Without loss of generality, we set $\xi = 0$ then in what follows we use notation $SN\left(\sigma^2, \delta\right)$ for simplicity. This defines the Poisson skew-normal (PSN) regression model as follows, where $'$ denotes vector transpose.

Definition 2 *Let for subject $i = 1, 2, \cdots, n$ and at time $t = 1, 2, \cdots, T$ the count variable $Y_{it}|u_{it} \overset{ind}{\sim} Pois(\exp(\theta_{it}))$, where $\theta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}$ and $u_{it} \overset{iid}{\sim} SN\left(\sigma^2, \delta\right)$. Then, the pmf of Y_{it} is of the form*

$$f_{PSN}\left(y_{it}|\boldsymbol{\beta}, \sigma^2, \delta\right) = \int_{-\infty}^{\infty} f_{Pois}\left(y_{it}|u_{it}, \boldsymbol{\beta}\right) g_{SN}\left(u_{it}|\sigma^2, \delta\right) du_{it}, \tag{2}$$

where $f_{Pois}\left(y_{it}|u_{it}, \boldsymbol{\beta}\right)$ is the conditional pmf of Poisson given u_{it}. We denote $Y_{it} \overset{ind}{\sim} PSN\left(\boldsymbol{\beta}, \sigma^2, \delta\right)$.

Clearly, the PN model is a special case of (2) when $\delta = 0$. Let $\mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta})$. By conducting algebraic operations, some properties of (2) are shown below in which any clear proof is omitted.

Properties 2

i. The mean and variance of Y_{it} are shown to be $E(Y_{it}) = \mu_{it}\mu_w$ and $\text{var}(Y_{it}) = \mu_{it}\left(\mu_w + \mu_{it}\sigma_w^2\right)$ so that the heterogeneity factor is $\left(\mu_w + \mu_{it}\sigma_w^2\right)/\mu_w$.
ii. The PSN is unimodal.

Proof. Since the skew-normal is unimodal thus the marginal mixed Poisson is also unimodal (Holgate, 1970).

iii. The PSN tends to $Pois(\mu_{it})$ as both σ^2 and δ tend to zero.

Proof. The normal mixing prior is regained for u_{it} as $\delta \to 0$. Then, using the transformation $v_{it} = exp(u_{it}/\sigma)$, the pmf (2) can be written as $E_{v_{it}}\{f_{Pois}(y_{it}|\mu_{it}v_{it}^{\sigma})\}$,

where v_{it} is log-normally distributed with $f_{v_{it}}(v_{it}) = 2\varphi(log(v_{it}))$, $v_{it} \in R^+$. Taking limit as $\sigma^2 \to 0^+$ this expectation becomes $f_{Pois}(y_{it}|\mu_{it})$.

iv. For fixed μ_{it} and non-zero δ, let $\sigma^2 \to 0^+$. Then (2) tends to a mixed Poisson density with a truncated normal mixing density supported on the left-bounded interval $(0,\infty)$, for $\delta > 0$, and on the right bounded interval $(-\infty,0)$, for $\delta < 0$.

Proof. We first derive the limiting case of $g_{SN}\left(u_{it}|\sigma^2,\delta\right)$ as $\sigma^2 \to 0^+$. It is easy to show that the density tends to $g_1(u_{it}|\delta) = 2\varphi\left(u_{it}|0,\delta^2\right)$ for $\text{sign}(u_{it}\delta) = 1$, and to 0 otherwise, where $sign(\cdot)$ denotes the sign function. Thus, the random variable Y_{it} follows a mixed Poisson (2) with the mixing prior g_1.

v. For fixed μ_{it} and σ^2 the probability $Pr(Y_{it} = 0)$ is a decreasing function of δ.

Proof. By setting $y_{it} = 0$ in (2) and taking the transformation $Z_{it} = \log(w_{it})$, the first derivative of the probability of zero is given by

$$\frac{\partial}{\partial\delta} f_{PSN}\left(0|\boldsymbol{\beta},\sigma^2,\delta\right) = \frac{\sigma^2}{\sqrt{\sigma^2+\delta^2}} E_{Z_{it}}\left(Z_{it}e^{-\mu_{it}e^{Z_{it}}}\right),$$

where $Z_{it} \sim N\left(0,\sigma^2\right)$. The involved expectation is shown to be negative. Then, after some manipulation twice of this expectation turns into $E\left(|Z_{it}|e^{-\mu_{it}e^{|Z_{it}|}}\right) - E\left(|Z_{it}|e^{-\mu_{it}e^{-|Z_{it}|}}\right)$. This expression is negative since the first expectation is less than the second one. This property is also illustrated by Figure 1.

vi. The probability of Y_{it} being zero is greater than the corresponding probability for a Poisson distribution with the same mean $\mu_{it}\mu_w$.
Proof. We have $f_{PSN}\left(0|\boldsymbol{\beta},\sigma^2,\delta\right) = E_{w_{it}}\left(e^{-\mu_{it}w_{it}}\right)$ and by using the Jensen's inequality, this becomes greater than $e^{-\mu_{it}\mu_w} = f_{Pois}(0|\mu_{it}\mu_w)$.

Figure 1 indicates the pmf of PSN for $\mu_{it} \equiv \mu = 3$, $\sigma^2 = 1$ and $\delta = -2,-1,0,1,2$. It is seen that the PSN is skewed right. Also, the tail of the PSN distribution is longer than

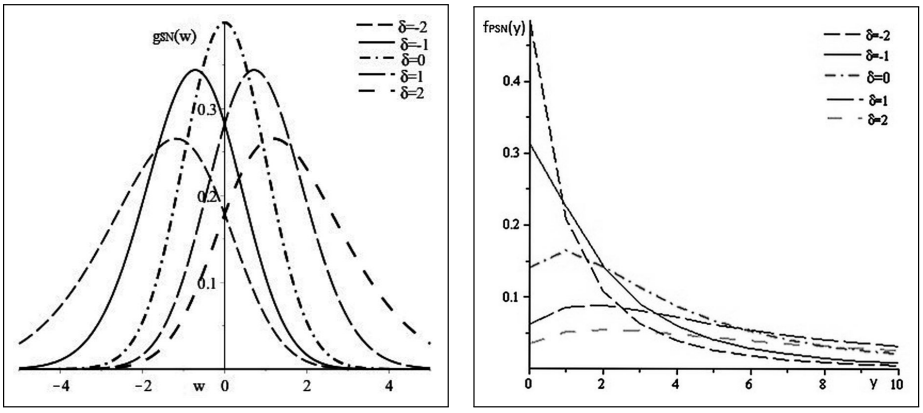


Figure 1: *(a) Probability density functions of the skew-normal distribution (b) Probability mass functions of the PSN distribution.*

the tail of the PN distribution for positive values of δ, while is shorter for negative values of δ. Furthermore, the probability of zero counts increases as δ decreases.
The dispersion index, defined by the ratio of the variance to the mean, is given by

$$DI_{it}\left(\mu_{it},\sigma^2,\delta\right)=\frac{\text{var}\left(Y_{it}\right)}{E\left(Y_{it}\right)}=1+\mu_{it}e^{\frac{1}{2}\left(\sigma^2+\delta^2\right)}\left\{\frac{\Phi\left(2\delta\right)e^{\left(\sigma^2+\delta^2\right)}-2\Phi^2\left(\delta\right)}{\Phi\left(\delta\right)}\right\}. \qquad (3)$$

This indicates that $DI_{it}>1$, with strict inequality if the mixing distribution is non-degenerate, i.e., the mixing strategy can deal with additional variation present in count data. If $\delta=0$ then (3) reduces to the DI of the PN regression model, denoted by $DI_{it}\left(\mu_{it},\sigma^2,0\right)$. The difference between the dispersion index of two densities, $DDI_{it}\left(\mu_{it},\sigma^2,\delta\right)=DI_{it}\left(\mu_{it},\sigma^2,\delta\right)-DI_{it}\left(\mu_{it},\sigma^2,0\right)$, is shown in Figure **2**. Negative and positive values of DDI_{it} show an advantage over the PN model. This indicates that the proposed model is more flexible than the PN model for dealing with over-dispersion in count data. Specifically, we set $\mu_{it}\equiv\mu=3$, $\sigma^2=1$ and $\delta\in(-1,1)$ that gives $DDI\left(\delta\right)\in(-2.446,47.769)$. Figure 2 illustrates that the PSN dispersion index is more than the PN dispersion index provided that $\delta>0$ while the difference $DDI\left(\delta\right)$ is negative for $\delta<0$. The differences increase as δ increases. We can also show by graphical techniques that if $\delta<0$ then the quantity $DDI\left(\delta\right)$ is positive over $\sigma^2\in\left(0,\sigma_0^2\right)$ for some small σ_0^2, whereas it is always negative over an interval $\sigma^2\in\left(\sigma_0^2,\infty\right)$. For any fixed δ the absolute value of $DDI(\delta)$ increases as σ^2 increases. Also, $DDI(\delta)<0$ for $\delta<0$, while $DDI(\delta)>0$ for $\delta>0$. These graphics are not shown here to save space.

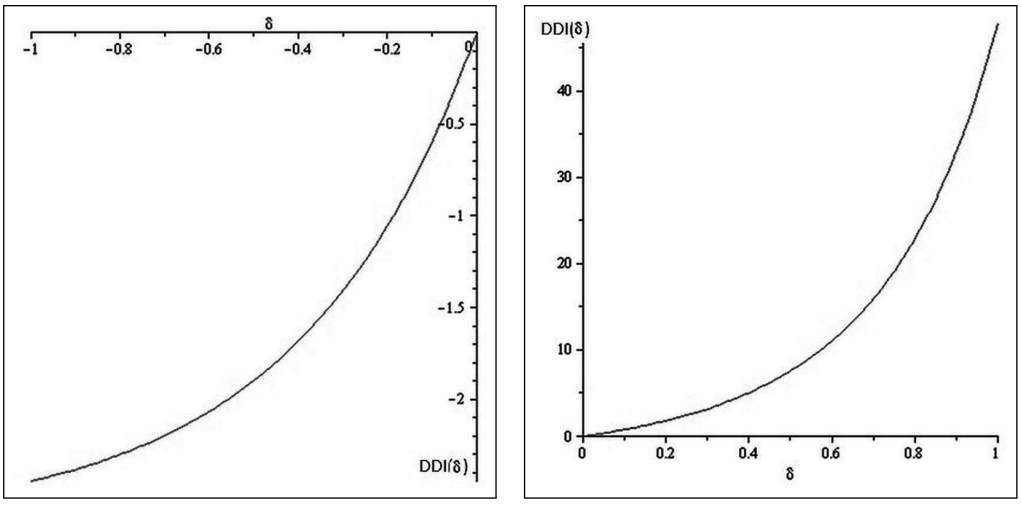


Figure 2: *Difference between the DIs for (a) negative and (b) positive values of* δ.

3 The proposed multivariate strategy for correlated count data

In longitudinal studies to count data, responses of each subject over time are usually correlated due to the existence of subject heterogeneity effects. Also, there may be evidence of over-dispersion in the structure of count responses. In many applications, there are situations where over-dispersion and the correlation between repeated outcomes can simultaneously occur. Here, we propose the multivariate skew-normal mixing prior distribution in the mean structure of mixed Poisson models to make a more adaptable analysis of correlated count responses. This strategy proceeds within the context of Bayesian hierarchical modelling together with several constructed specifications to fit related regression models. The specification of these proposed Poisson multivariate skew-normal (PMSN) models relies mostly on making different assumptions for the underlying multivariate skew-normal mixing priors.

3.1 The multivariate skew-normal mixing priors

Definition 3 *A T-dimensional random vector $\mathbf{u}$ follows the multivariate skew-normal distribution with location vector $\boldsymbol{\xi} \in \mathbb{R}^T$, positive-definite scale matrix $\mathbf{V}$, and skewness vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_T)' \in \mathbb{R}^T$, if its pdf is of the form*

$$f(\mathbf{u}_i|\boldsymbol{\xi}, \mathbf{V}, \boldsymbol{\delta}) = 2\varphi_T\left(\mathbf{u}_i|\boldsymbol{\xi}, V + \boldsymbol{\delta}\boldsymbol{\delta}'\right)\Phi\left(\frac{\boldsymbol{\delta}'\mathbf{V}^{-1}(\mathbf{u}_i - \boldsymbol{\xi})}{\sqrt{1 + \boldsymbol{\delta}'\mathbf{V}^{-1}\boldsymbol{\delta}}}\right), \tag{4}$$

where $\varphi_T(.)$ is the pdf of T-variate normal and $\Phi(\cdot)$ is the standard normal cdf. We denote $\mathbf{u}_i \sim SN_T(\boldsymbol{\xi}, \mathbf{V}, \boldsymbol{\delta})$.

The density function (4) defines an attractive alternative to the multivariate skew-normal distribution introduced previously by Sahu et al. (2003) since instead of the evaluation of complex function $\Phi_T(\cdot)$, one needs only to compute one dimensional integral $\Phi(\cdot)$. The Poisson multivariate normal (PMN) model is a special case of (4) when $\boldsymbol{\delta} = \mathbf{0}$.

Properties 3 *The following properties hold for $\mathbf{u}_i \sim SN_T(\boldsymbol{\xi}, \mathbf{V}, \boldsymbol{\delta})$:*

i. The hierarchical representation is given by

$$\mathbf{u}_i|Z_i = z_i \overset{ind}{\sim} N_T(\boldsymbol{\xi} + \boldsymbol{\delta}z_i, \mathbf{V}) \text{ with } Z_i \overset{iid}{\sim} HN(0,1), \tag{5}$$

Thus, the mean vector and covariance matrix of $\mathbf{u}_i$ can be derived relatively easy. We obtain $E(\mathbf{u}_i) = \boldsymbol{\xi} + \boldsymbol{\delta}\sqrt{\frac{2}{\pi}}$, and $\text{var}(\mathbf{u}_i) = \mathbf{V} + \left(1 - \frac{2}{\pi}\right)\boldsymbol{\delta}\boldsymbol{\delta}'$.

ii. For any vector $\mathbf{r} = (r_1, \cdots, r_T)' \in \mathbb{R}^T$ the MGF is found to be

$$E\left(e^{\mathbf{r}'\mathbf{u}_i}\right) = 2\Phi(\mathbf{r}'\boldsymbol{\delta}) \exp\left\{\mathbf{r}'\boldsymbol{\xi} + \frac{1}{2}\mathbf{r}'\mathbf{V}\mathbf{r} + \left(\boldsymbol{\delta}'\mathbf{r}\right)^2\right\}. \tag{6}$$

Now, let $w_{it} = \exp(u_{it})$ be an element of the vector $w_i = (w_{i1}, \cdots, w_{iT})'$. Equation (6) is equivalent to $E\left(\prod_{t=1}^{T} w_{it}^{r_t}\right)$ which shows that all moments of w_{it}, including $E(w_i) = \boldsymbol{\mu}_w = \left(\mu_{w_{i1}}, \cdots, \mu_{w_{iT}}\right)'$ and $\text{var}(w_i) = D_w$, can be found easily. Specifically, putting the t-th element of $\mathbf{r}$ equal to one and zero otherwise, gives $\mu_{w_{it}}$, and when $r_t = r_s = 1$ and 0 otherwise, $E(w_{it}w_{is})$ is attained. In fact, we derive

$$\begin{aligned} \mu_{w_{it}} &= 2\Phi(\delta_t)\, e^{\frac{1}{2}\left(\delta_t^2 + \sigma_{tt}\right)}, \\ \sigma_{w_{ts}} &= 2e^{\frac{1}{2}\left(\delta_t^2 + \delta_s^2 + \sigma_{tt} + \sigma_{ss}\right)} \left\{ e^{\delta_t\delta_s + \sigma_{ts}} \Phi(\delta_t + \delta_s) - 2\Phi(\delta_t)\Phi(\delta_s) \right\}, \end{aligned} \tag{7}$$

where the $\sigma_{w_{ts}}$ and σ_{ts} are, respectively, elements of $\mathbf{D}_w$ and $\mathbf{V}$.

iii. Let $c = \mathbf{a}'\mathbf{u}_i$ for any $\mathbf{a} \in \mathbb{R}^T$ then c follows the univariate skew-normal distribution, i.e. $c \sim SN(\mathbf{a}'\boldsymbol{\xi}, \mathbf{a}'\mathbf{V}\mathbf{a}, \mathbf{a}'\boldsymbol{\delta})$.

Without loss of generality, in what follows we set $\boldsymbol{\xi} = \mathbf{0}$ and denote $\mathbf{u}_i \sim SN_T(\mathbf{V}, \boldsymbol{\delta})$ for simplicity. In model multivariate skew-normal specified by (4) no specific form of V and $\boldsymbol{\delta}$ is introduced in the data analysis process. It is mostly advisable in practice to explore possible causes of heterogeneity by allowing some specific forms for the u_{it}'s. Without having any knowledge on the source of heterogeneity, a priori justification is to allow u_{it}'s being into the one-way random effects framework. More specifically, let the $\mathbf{u}_i$ be of the familiar form $\mathbf{u}_i = \alpha_i \mathbf{1}_T + \boldsymbol{\varepsilon}_i$, where $\mathbf{1}_T$ denotes a unit vector of order T, the α_i represent the heterogeneity effects and the $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iT})'$ denote the residual terms that may reflect time-varying effects such as the effect of unobserved omitted covariates. In this setting, we specify the following types of the multivariate skew-normal distribution.

Remark 1 *For the above specified multivariate skew-normal model, let* $\alpha_i \overset{iid}{\sim} N\left(0, \sigma_\alpha^2\right)$ *and* $\boldsymbol{\varepsilon}_i \overset{iid}{\sim} SN_T(\mathbf{V}_\varepsilon, \boldsymbol{\delta})$ *be all mutually independent. Then* $\mathbf{u}_i \overset{iid}{\sim} SN_T(\mathbf{D}, \boldsymbol{\delta})$ *where* $\mathbf{D} = \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T' + \mathbf{V}_\varepsilon$.

For the case with $\mathbf{V}_\varepsilon$ diagonal, we obtain

$$\text{corr}(w_{it}, w_{is}) = \frac{e^{\delta_t\delta_s + \sigma_\alpha^2}\Phi(\delta_t + \delta_s) - 2\Phi(\delta_t)\Phi(\delta_s)}{\sqrt{e^{\sigma_{tt} + \delta_t^2 + \sigma_\alpha^2}\Phi(2\delta_t) - 2\Phi^2(\delta_t)}\sqrt{e^{\sigma_{ss} + \delta_s^2 + \sigma_\alpha^2}\Phi(2\delta_s) - 2\Phi^2(\delta_s)}}, \tag{8}$$

for any $t \neq s$. Note that, the correlation coefficient (8) may take negative or positive values in the interval (-1,1).

Remark 2 *For the familiar form* $\mathbf{u}_i = \alpha_i \mathbf{1}_T + \boldsymbol{\varepsilon}_i$, *let* $\alpha_i \overset{iid}{\sim} SN\left(\sigma_\alpha^2, \delta\right)$ *and* $\boldsymbol{\varepsilon}_i \overset{iid}{\sim} N_T\left(\mathbf{0}, \mathbf{V}_\varepsilon\right)$ *be all mutually independent. Then* $\mathbf{u}_i \overset{iid}{\sim} SN_T\left(\mathbf{D}, \delta \mathbf{1}_T\right)$.

The correlation between w_{it} and w_{is} is a special case of (8) when δ_t and δ_s are replaced by constant δ for all t, s.

3.2 The Poisson multivariate skew-normal model

Let Y_{it} be the response variable and u_{it} be the corresponding heterogeneity effect of subject i at time period t for $i = 1, 2, \cdots, n$ and $t = 1, 2, \cdots, T$. The scheme of a PSN regression model in (2) allows for over-dispersion in the Poisson model but without taking into account the correlation among events. A common way to deal with this issue is to allow the vector $\mathbf{u}_i = (u_{i1}, \cdots, u_{iT})'$ to follow a multivariate distribution with correlation amongst $u_{i1}, \cdots, u_{iT}$, and consequently induce correlated $Y_{i1}, \cdots, Y_{iT}$. A frequent assumption is multivariate normality of the $\mathbf{u}_i$. An alternative is to utilize a multivariate skew-normal distribution. Several versions of the multivariate skew-normal distribution, originally introduced by Azzalini and Dalla Valle (1996), have appeared in the literature. We present below a slight alteration of this distribution and provide its main properties that are related to our current work.

Definition 4 *Let the response vectors* $\mathbf{Y}_i = \{Y_{it}\}$ *of order T be independent for subjects* $i = 1, \cdots, n$ *and each* Y_{it} *conditioned on the effect* $\mathbf{u}_i$ *follows Poisson with the conditional mean* $exp(\theta_{it})$, *where* $\theta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}$ *and* $\mathbf{u}_i = \{u_{it}\} \overset{iid}{\sim} SN_T\left(\mathbf{V}, \boldsymbol{\delta}\right)$. *Then the marginal pmf of* $\mathbf{Y}_i$ *is given by*

$$f_{PMSN}\left(\mathbf{y}_i | \boldsymbol{\beta}, \mathbf{V}, \boldsymbol{\delta}\right) = \int_{\mathbb{R}^T} \prod_{t=1}^{T} f_{Pois}\left(y_{it} | u_{it}, \boldsymbol{\beta}\right) g_{MSN}\left(\mathbf{u}_i | \mathbf{V}, \boldsymbol{\delta}\right) d\mathbf{u}_i, \tag{9}$$

where $g_{MSN}\left(\mathbf{u}_i | \mathbf{V}, \boldsymbol{\delta}\right)$ *denotes the multivariate skew-normal density function for the i-th subject. We denote* (9) *as model PMSN*1.

The solution of (9) is not generally available in closed form. Thus, an MCMC scheme is implemented later to make statistical inferences. Furthermore, through standard calculation (see the supplementary Appendix A) we can straightforwardly show that

$$E\left(\mathbf{Y}_i\right) = \mathbf{M}_i \boldsymbol{\mu}_w, \text{ and } \operatorname{var}\left(\mathbf{Y}_i\right) = \mathbf{M}_i \mathbf{D}_w \mathbf{M}_i + \mathbf{M}_i \mathbf{M}_w, \tag{10}$$

where $\mathbf{M}_w$ and $\mathbf{M}_i$ are diagonal matrices with the elements $\mu_{w_{it}}$ and μ_{it} for $t = 1, 2, \cdots, T$, respectively. The corresponding correlation coefficients between counts Y_{it} and Y_{is} are

given by

$$\operatorname{corr}(Y_{it}, Y_{is}) = \operatorname{corr}(w_{it}, w_{is}) \sqrt{\frac{\mu_{it}}{\mu_{it} + \frac{\mu_{w_{it}}}{\sigma_{w_{tt}}}}} \sqrt{\frac{\mu_{is}}{\mu_{is} + \frac{\mu_{w_{is}}}{\sigma_{w_{ss}}}}}, \tag{11}$$

for all i, t, and s. Equation (11) shows that the two correlations $\operatorname{corr}(Y_{it}, Y_{is})$ and $\operatorname{corr}(w_{it}, w_{is})$ have the same sign and that $|\operatorname{corr}(Y_{it}, Y_{is})| < |\operatorname{corr}(w_{it}, w_{is})|$. Also, negative and positive correlations are allowed by using these mixed models. This fact gives an advantage over other multivariate models for discrete outcomes such as multinomial or negative multinomial models that allow only positive correlation. We specify below two types of the PMSN1 model.

Definition 5 *Let* $\mathbf{u}_i \overset{iid}{\sim} SN_T(\mathbf{D}, \boldsymbol{\delta})$. *We denote the corresponding model as PMSN2.*

The mean vector and covariance matrix of $\mathbf{Y}_i$ are derived to be particular cases of (10) and setting (7) in which the scalar σ_{ts}, for $t, s = 1, 2, \cdots, T$, turns into $\sigma_{ts} + \sigma_{\alpha}^2$, i.e. elements of $\mathbf{D}$. In this model, the corresponding correlation coefficient may take negative or positive values.

Definition 6 *Let* $\mathbf{u}_i \overset{iid}{\sim} SN_T(\mathbf{D}, \delta \mathbf{1}_T)$. *We denote the corresponding model as PMSN3.*

By utilizing the correlation between w_{it} and w_{is}, the resultant equation is always positive showing that PMSN3 permits only positive correlation between events. For a model with only constant term (no explanatory variable) Equations (10) can be simplified as $\operatorname{var}(Y_{it}) = c\mu_v\mu_\psi + c^2\left(\sigma_\psi^2\sigma_v^2 + \mu_v^2\sigma_\psi^2 + \mu_\psi^2\sigma_v^2\right)$, $\operatorname{cov}(Y_{it}, Y_{is}) = c^2\mu_v^2\sigma_\psi^2$, $t \neq s$ where $c = \exp(\beta_0)$ and parameters μ_v, σ_v^2, μ_ψ and σ_ψ^2 denote, correspondingly,means and variances of $v_{it} = \exp(\varepsilon_{it})$ and $\psi_i = \exp(\alpha_i)$ in $\mathbf{u}_i = \alpha_i \mathbf{1}_T + \boldsymbol{\varepsilon}_i$. It follows that

$$\operatorname{corr}(Y_{it}, Y_{is}) = \frac{c\mu_v^2\sigma_\psi^2}{1 + c\left(\sigma_\psi^2\sigma_v^2 + \mu_v^2\sigma_\psi^2 + \mu_\psi^2\sigma_v^2\right)}, \; t \neq s. \tag{12}$$

If the estimate of (12) is statistically significant then the PMSN model fits better to the data set than the standard Poisson regression model.

3.3 An alternative to deal with the identification issue

In the literature of mixed Poisson models the identification is usually addressed by allowing a restriction to the estimation process in order to make estimable the model parameters. To clarify this, let the count Y_{it}, for subject $i = 1, 2, \cdots, n$ and at time $t = 1, 2, \cdots, T$, follows the PSN distribution in (2), where $\log(E(Y_{it})) = \mathbf{x}_{it}'\boldsymbol{\beta} + \log(\mu_w)$. A common approach used by many researchers (e.g. see Balakrishnan and Peng, 2006) is to reparameterize the mixing distribution such that $\mu_w = 1$ to ensure that the loga-

rithm of the marginal expectation of counts is $\mathbf{x}'_{it}\boldsymbol{\beta}$. This is equivalent to solving the nonlinear equation $2\Phi(\delta) = \exp\left\{-0.5\left(\delta^2+\sigma^2\right)\right\}$ for δ. However, this method does not work well when the expectation of exponentiated unobserved heterogeneity has a complex structure. Also, it may cause difficulties in the process of optimization routines for the reparameterized model. Thus, we use an alternative trick by setting the regression parameter β_0 to be equal to $\log(\mu_w)$. A similar trick can be done for models PMSN1–PMSN3 by setting $\beta_{0t} = \log(\mu_{w_t})$ since μ_{w_t} depends on time t.

4 The computational scheme

This section develops an operational MCMC scheme for the Bayesian analysis of the proposed regression models. We utilize the Bayesian data-augmentation method (e.g., Albert and Chib, 1993), which lets us generate the heterogeneity effects along with other known quantities in the simulation process. We use the hierarchical representation of all models to write down the joint likelihood of responses and heterogeneity effects. This representation is quite useful to estimate parameters by using the MCMC technique. To complete the model specifications from a Bayesian perspective, we assume that all parameters are independent. Then, we assign conditionally semi-conjugate priors to these parameters. This choice simplifies computations since the complete conditional posteriors involved in the Gibbs sampler are mostly closed forms of known distributions and hence easy for simulation. It this section, we let $Z_{it} \overset{iid}{\sim} HN(0,1)$ and $Z_i \overset{iid}{\sim} HN(0,1)$.

4.1 Bayesian computation for independent data

To fit the PSN model, we use the data augmentation to θ_{it} based on the hierarchical representation of the skew-normal distribution given in Properties 1 (ii). The related hierarchical form becomes

$$\begin{aligned} &Y_{it}|\theta_{it} \overset{ind}{\sim} Pois(\exp(\theta_{it})), \\ &\theta_{it}|z_{it},\boldsymbol{\beta},\sigma^2,\delta \overset{ind}{\sim} N\left(\mathbf{x}'_{it}\boldsymbol{\beta}+\delta z_{it},\sigma^2\right), \end{aligned} \tag{13}$$

for subject $i = 1,2,\cdots,n$ and at time $t = 1,2,\cdots,T$. By adopting all parameters to be independent, we assign the priors $\boldsymbol{\beta} \sim N_k\left(\boldsymbol{\beta}_0, \mathbf{V}_{\boldsymbol{\beta}}\right)$, $\delta \sim N\left(\delta_0, \sigma^2_\delta\right)$, and an inverse-Gamma, $IG(\nu_0,\nu_0)$, for σ^2, where all hyperparameters are known. The joint posterior density of $\boldsymbol{\beta}$, σ^2 **,** δ, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_n)'$ and $\mathbf{z} = (\mathbf{z}_1,\cdots,\mathbf{z}_n)'$ with $\boldsymbol{\theta}_i = (\theta_{i1},\cdots,\theta_{iT})'$ and $\mathbf{z}_i = (z_{i1},\cdots,z_{iT})'$ is then given by

$$\begin{aligned} \pi\left(\boldsymbol{\beta},\sigma^2,\delta,\boldsymbol{\theta},\mathbf{z}\right) &\propto \prod_{i=1}^{n}\prod_{t=1}^{T} f_{Pois}\left(y_{it}|\theta_i\right)\varphi\left(\theta_{it}|\mathbf{x}'_{it}\boldsymbol{\beta}+\delta z_{it},\sigma^2\right) \\ \varphi(z_{it}|0,1)\,I(z_{it}>0) &\times \varphi_k\left(\boldsymbol{\beta}|\boldsymbol{\beta}_0,\mathbf{V}_{\boldsymbol{\beta}}\right)\varphi\left(\delta|\delta_0,\sigma^2_\delta\right) f_{IG}\left(\sigma^2|\nu_0,\nu_0\right). \end{aligned} \tag{14}$$

The marginal posterior is derived by integrating out $\boldsymbol{\theta}$ and $\mathbf{z}$ from (14). This posterior is analytically intractable and the solution requires implementing advanced numerical integration techniques or utilizing the MCMC procedures, such as Gibbs sampling. The Gibbs sampling algorithm simulates iteratively from the complete conditional posterior distribution of each unknown stochastic parameter, or quantity, conditioned on the remaining parameters and unknown quantities. The complete conditional posterior distributions are given in the supplementary Appendix B.

4.2 Bayesian computation for the correlated data

Here, we use the following hierarchical representation of the defined models. The Gibbs sampling to fit model PMSN1 is implemented as follows.

The PMSN1 model. Consider the hierarchical form

$$\begin{aligned} &Y_{it}|\theta_{it} \overset{iid}{\sim} Pois\left(\exp(\theta_{it})\right), \\ &\boldsymbol{\theta}_i|z_i,\boldsymbol{\beta},\mathbf{V},\boldsymbol{\delta} \overset{ind}{\sim} N_T\left(\mathbf{X}_i\boldsymbol{\beta}+\boldsymbol{\delta}z_i,\mathbf{V}\right), \end{aligned} \tag{15}$$

for subject $i=1,2,\cdots,n$ and at time $t=1,2,\cdots,T$. Assuming the multivariate normal prior for $\boldsymbol{\beta}$, the inverse-Wishart $IW_T\left(\boldsymbol{\Omega},m\right)$ for matrix $\mathbf{V}$, and $N_T\left(\boldsymbol{\delta}_0,\mathbf{V}_{\boldsymbol{\delta}}\right)$ for the vector of skewness parameters $\boldsymbol{\delta}$, where all hyper-parameters are assumed to be known, we derive the related complete conditional posteriors as given in the supplementary Appendix B. The specification of models PMSN2 and PMSN3 are given below by using the multivariate skew-normal mixing prior.

The PMSN2 model. The hierarchical form of PMSN2 is

$$\begin{aligned} &Y_{it}|\theta_{it} \overset{ind}{\sim} Pois\left(\exp(\theta_{it})\right), \\ &\boldsymbol{\theta}_i|\alpha_i,z_i,\boldsymbol{\beta},\mathbf{V}_\varepsilon,\boldsymbol{\delta} \overset{ind}{\sim} N_T\left(\mathbf{X}_i\boldsymbol{\beta}+\alpha_i\mathbf{1}_T+\boldsymbol{\delta}z_i,\mathbf{V}_\varepsilon\right), \\ &\alpha_i \overset{iid}{\sim} N\left(0,\sigma_\alpha^2\right). \end{aligned} \tag{16}$$

The PMSN3 model. The hierarchical form of PMSN3 is

$$\begin{aligned} &Y_{it}|\theta_{it} \overset{ind}{\sim} Pois\left(\exp(\theta_{it})\right), \\ &\boldsymbol{\theta}_i|\alpha_i,z_i,\boldsymbol{\beta},\mathbf{V}_\varepsilon,\delta \overset{ind}{\sim} N_T\left(\mathbf{X}_i\boldsymbol{\beta}+\alpha_i\mathbf{1}_T,\mathbf{V}_\varepsilon\right), \\ &\alpha_i|z_i \overset{ind}{\sim} N\left(\delta z_i,\sigma_\alpha^2\right). \end{aligned} \tag{17}$$

The Bayesian computational details of mixed Poisson models PMSN2 and PMSN3, including priors and complete conditional posteriors, are given in supplementary Appendix B. All complete conditional posteriors, except for $\boldsymbol{\theta}$, appear in closed forms of known distributions and thus random samples can easily be generated. However,

drawing samples from the posterior of $\boldsymbol{\theta}$ maybe done by the accept-reject algorithm (Gilks and Wild, 1992) or by the Metropolis-Hastings algorithm within the Gibbs sampler (Chib and Greenberg, 1995). Thus, the Gibbs sampler proceeds by simulating a sequence of samples from the complete conditional posteriors. The sampler simulates iteratively from these posteriors by running a sufficient burn-in period until convergence to stationary distributions occurs. Then, the average of samples for each parameter is used as its Bayes estimate. Convergence is monitored via MCMC chain histories, Gelman-Rubin diagnostic, autocorrelation, and density plots.

5 Comparative studies using simulation

We conduct two simulation studies to highlight the usefulness of proposed models. Specifically, we design Monte Carlo experiments to underline the important role of the skewness parameter and the structure of covariance matrices. We also make comparisons between competing models. To implement the Gibbs sampler, the following independent priors are adopted: $N(0,100)$ for the regression coefficients as well as for δ, Uniform$(-1,1)$ for ρ, and Inverse-Gamma$(0.01,0.01)$ for the variance components. Using the OpenBUGs software version 3.2.3, we run 10,000 samples after removing 5,000 burn-in until the convergence occurs. There was no evidence of lack of convergence based on examinations of histories, Gelman-Rubin diagnostic, kernel density, and autocorrelation plots. Also, by using various values of hyperparameters, we obtained similar results, which implies that posterior estimates are not sensitive to the prior in this Bayesian analysis.

a. The simulated model is PMSN1: We generated 1,000 independent Monte Carlo data sets from model PMSN1 with $n = 100$ sample size. Consider the longitudinal data model

$$Y_{it}|\theta_{it} \overset{ind}{\sim} Pois\left(\exp(\theta_{it})\right) \text{ with } \theta_{it} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{it2} + u_{it}, \tag{18}$$

for subject $i = 1,2,\cdots,100$ and at time $t = 1,2,\cdots,5$. Random counts are generated according to (9), where $\mathbf{u}_i \overset{iid}{\sim} SN_2(\mathbf{V},\boldsymbol{\delta})$ with $\boldsymbol{\delta}' = \delta \mathbf{1}_5'$ and $\mathbf{V} = \left\{\sigma^2 \rho^{|t-s|}\right\}$ for $t,s = 1,2,\cdots,5$. The time-constant covariate X_{i1} is generated by $Bernoulli(0.5)$ and the time-varying covariate X_{it2} by $N(0,1)$. For all experiments, θ_{it} was computed by setting $\beta_1 = -1$ and $\beta_2 = 1$. We set $\rho = 0.5$, $\sigma^2 = 0.36$, and $\delta = -0.8$, 0, 0.8. Taking into account the identification issue, we obtain $\beta_0 = -0.36, 0.18$ and 0.96, respectively. Results are reported in Tables 1 and 2 along with the fitted standard Poisson model for comparison. Biases and the mean squared error (MSE $\times 10$) of estimates are computed. Smaller values of the MSE indicate a better fit.

In each generation, the variance of Y differs considerably from the mean of Y, unlike the conventional Poisson density. Figure 3 illustrates this feature for $\delta = 0.8$ and the first 100 generations. This shows strong evidence of over-dispersion. Thus, fitting mixed Poisson models may be more appropriate to this data set.

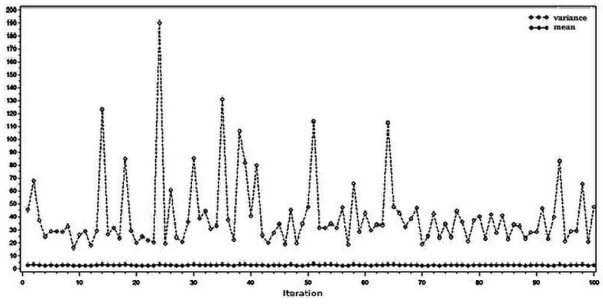


Figure 3: *The first 100 generations of model PMSN1 with* $\delta = 0.8$.

Therefore, we fit the hypothetical models PMN with $\mathbf{V} = \left\{\sigma^2 \rho^{|t-s|}\right\}$ and PMSN2 with $\mathbf{D} = \sigma_\alpha^2 \mathbf{1}_5 \mathbf{1}_5' + \sigma^2 \mathbf{I}_5$, where $\boldsymbol{\delta} = \delta \mathbf{1}_5$.

Table 1: *Biases and MSEs(*$\times 10$*) for the proposed models.*

		Poisson		PMN		PMSN1		PMSN2	
δ		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
-0.8	β_0	-0.017	3.404	0.582	2.629	-0.021	0.005	0.038	0.016
	β_1	-0.117	2.138	-0.512	2.036	-0.013	0.004	-0.140	0.199
	β_2	-0.036	0.713	-0.155	0.160	-0.003	0.001	-0.031	0.010
	σ^2			0.085	0.031	-0.001	0.001	-0.035	0.013
	δ					-0.012	0.004	0.107	0.116
	ρ			0.156	0.244	-0.019	0.007		
0	β_0	0.018	0.904	0.002	0.001	$<$0.001	0.001	-0.076	0.059
	β_1	-0.012	0.202	-0.002	0.001	$<$0.001	0.001	-0.010	0.003
	β_2	-0.004	0.002	-0.001	0.001	$<$0.001	0.001	0.006	0.002
	σ^2			0.004	0.002	0.012	0.002	-0.095	0.090
	δ					-0.014	0.003	-0.042	0.019
	ρ			-0.050	0.029	-0.025	0.008		
0.8	β_0	-0.060	2.037	-0.478	2.284	-0.002	0.001	-0.174	0.304
	β_1	0.032	0.311	0.487	2.374	-0.007	0.002	0.123	0.153
	β_2	-0.018	0.210	0.107	0.115	-0.010	0.001	0.035	0.013
	σ^2			0.287	0.825	-0.040	0.017	0.168	0.284
	δ					-0.017	0.004	-0.233	0.545
	ρ			-0.080	0.064	-0.004	0.006		

Results, after the convergence is achieved, are reported in Table *1*. Note that, we let $\boldsymbol{\delta} = \delta \mathbf{1}_5$ which implies equivalence of PMSN2 and PMSN3 models. Also, the concern was to illustrate the impact of ignoring dependency between the u_{it}'s for $t = 1, 2, \cdots, T$. Thus, PSN was not fitted. For $\delta = 0$, the PMSN1 performs as well as the PMN model. This finding shows that the PMSN1 is a flexible model since it can cover either symmetric or asymmetric data, depending on the values of its skewness parameter. For $\delta = -0.8, 0.8$ and based on MSEs, the PMSN models, PMSN1 and PMSN2, are bet-

ter fitted than the conventional Poisson and PMN models. This finding does suggest the importance of identifying the correlation of the heterogeneity effects. To make a further comparison, we compute the relative efficiency $r = MSE_M/MSE_{PMSN1}$, where M denotes the competitive regression model. Efficiency values, shown in Table **2**, are remarkably greater than 1, illustrating that the PMSN1 estimates are efficient compared to the parameter estimates in the hypothetical regression models.

Table 2: *Relative efficiencies of estimates in the longitudinal study.*

δ		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma^2}$	$\hat{\delta}$	$\hat{\rho}$
-0.8	Poisson	680.8	534.5	713.0			
	PMN	525.8	509.0	160.0	31.0		34.8
	PMSN2	3.2	49.7	10.0	13.0	29.0	
0	Poisson	904.0	202.0	2.0			
	PMN	1.0	1.0	1.0	2.0		3.6
	PMSN2	59.0	3.0	2.0	45.0	6.3	
0.8	Poisson	2037.0	155.5	210.0			
	PMN	2284.0	1187.0	115.0	48.5		10.667
	PMSN2	304.0	76.5	13.0	16.7	136.2	

b. The simulated model is PMSN2: Here, a simulation study is conducted to distinguish the performance of the PMSN1 model with $\mathbf{V} = \left\{\sigma^2\rho^{t-s}\right\}$, for $t,s = 1,2$, when response values are generated according to the PMSN2 model. We sampled data from (9) where θ_{it} is given in (18) and $\mathbf{u}_i \overset{iid}{\sim} SN_2(\mathbf{D},\boldsymbol{\delta})$ with $\boldsymbol{\delta}' = (\delta_1,\delta_2)$ and $\mathbf{D} = \sigma_\alpha^2\mathbf{1}_T\mathbf{1}_T' + \mathbf{V}_\varepsilon$. Also, the covariates X_{i1} and X_{it2} are generated respectively from a $Bernoulli(0.5)$ and a standard normal distribution. We set $\beta_1 = -1$, $\beta_2 = 1$, $\sigma_\alpha^2 = 0.25$, $\delta = -4,1$, and the variance components $\sigma_{\varepsilon,1}^2 = \sigma_{\varepsilon,2}^2 = 0.25$ and $\sigma_{\varepsilon,12} = 0.75$ for $\mathbf{V}_\varepsilon$. Now, the correlation between observations is negative. Taking into account the identification issue, we obtain $\beta_0 = -0.65$ and 2, respectively. All parameters are estimated using the PMSN1 model. The posterior means (each with standard deviation) of estimates are obtained as $\tilde{\beta}_0 = -0.663(0.031), 2.063(0.033)$, $\tilde{\beta}_1 = -1.0090.030$, $\tilde{\beta}_2 = 1.008(0.016)$, $\tilde{\delta}_1 = -4.144(0.069)$, $\tilde{\delta}_2 = 0.987(0.026)$, $\tilde{\rho} = 0.507(0.012)$, and $\tilde{\sigma}^2 = 2.035(0.056)$. We observe that the bias of each regression coefficient and skewness parameter is small. Thus, the evidence again recommends that the regression model PMSN1 is appropriate to analyse the data.

6 Empirical studies

This section considers two examples taken from the literature that have been previously analysed by several authors. We fit the proposed models using the OpenBugs software. Priors for the regression coefficients and the skewness parameter were assumed to be

independent, each distributed normally with zero mean and 0.001 precision, variance components distributed as inverse-Gamma distribution with parameters both equal to 0.1. The model fitting process has carried out for 10,000 iterations after discarding the first 5,000 iterations to ensure us the convergence has occurred. There was no evidence of lack of convergence due to examinations of histories, Gelman-Rubin diagnostic, kernel density, and autocorrelation plots. For illustration, the supplementary Appendix B shows posterior plots of the PMSN2 and β_5 for the health reform data.

A popular model selection in the Bayesian framework is the deviance information criterion (DIC). However, the DIC in OpenBugs is based on the conditional likelihood given the random effects. To compare the fitted models marginally we alternatively compute the Akaike information criterion, $AIC(\boldsymbol{\theta}) = D(\boldsymbol{\theta}) + 2p$, and the Bayesian information criterion, $BIC(\boldsymbol{\theta}) = D(\boldsymbol{\theta}) + p\log(n)$, where the deviance $D(\boldsymbol{\theta}) = -2\log L(\boldsymbol{\theta})$, p and n denote the number of parameters and sample size, respectively. To estimate $D(\boldsymbol{\theta})$, we use the deviance evaluated at the Bayes estimates of parameters $\boldsymbol{\theta}$, where $L(\boldsymbol{\theta})$ is taken as the underlying marginal likelihood. Smaller values of these criteria indicate better fit.

6.1 Polio incidence

The Polio data set is taken from the US Centers for Disease Control and Prevention. The response variable is the monthly number of poliomyelitis cases (Y), over the years 1970 to 1983. The data were previously analysed by several researchers, such as Zeger (1988), Oh and Lim (2001), Davis and Wu (2009), Fokianos and Fried (2012) and Kang and Lee (2014) between others. We fit a similar model as given by Zeger (1988), and Oh and Lim (2001) by noting that the regression model is organized in terms of a re-centred version of time t, such that it can be easily convenient within the usual framework of the cross-sectional data model. The model includes an intercept, a time trend, and some trigonometric components at periods 6 and 12 months. Fitting a Poisson model, the ratio of deviance to degrees-of-freedom was 1.925, illustrating evidence of over-dispersion. Thus, the Poisson model is not suitable to fit the data. We now fit the PSN regression model, for $n = 1$, already specified in Section 2. Specifically, let $Y_t|u_t \overset{ind}{\sim} Pois(\exp(\theta_t))$ for $t = 1, \cdots, 168$, where

$$\begin{aligned} \theta_t &= \beta_0 + \beta_1 t^* \times 10^{-3} + \beta_2 \cos\left(\frac{2\pi t^*}{6}\right) + \beta_3 \sin\left(\frac{2\pi t^*}{6}\right) \\ &+ \beta_4 \cos\left(\frac{2\pi t^*}{12}\right) + \beta_5 \sin\left(\frac{2\pi t^*}{12}\right) + \beta_6 y_{t-1} + u_t, \end{aligned}$$

and $t^* = t - 73$ is used to locate the intercept term at January 1976 as in Zeger's analysis. We also analyse the polio incidence rates using the PN model; i.e. $u_t \overset{iid}{\sim} N\left(0, \sigma^2\right)$. Bayes estimates, standard deviations, 95% confidence intervals, and some information criteria for models comparison are given in Table 3.

Table 3: *Posterior summary statistics for parameters of fitted models.*

Model	Poisson	PN	PSN
	Est(s.d.) (95% CI)	Est(s.d.) (95% CI)	Est(s.d.) (95% CI)
β_0	0.046(0.088) (-0.131,0.215)	0.220(0.068) (0.108,0.374)	0.031(0.084) (-0.119,0.217)
β_1	-3.753(1.441) (-6.554,-0.9409)	-4.160(1.816) (-7.754,-0.663)	-3.508(1.956) (-7.378,0.299)
β_2	0.101(0.103) (-0.099,0.304)	0.130(0.129) (-0.119,0.386)	0.125(0.140) (-0.144,0.403)
β_3	-0.410(0.102) (-0.612,-0.211)	-0.348(0.127) (-0.596,-0.099)	-0.358(0.137) (-0.629,-0.090)
β_4	-0.181(0.098) (-0.376,0.010)	-0.125(0.125) (-0.369,0.122)	-0.140(0.134) (-0.398,0.126)
β_5	-0.464(0.0.111) (-0.686,-0.250)	-0.443(0.136) (-0.712,-0.185)	-0.457(0.147) (-0.750,-0.164)
β_6	0.092(0.025) (0.041,0.140)	0.059(0.038) (-0.019,0.131)	0.104(0.041) (0.018,0.181)
σ^2		0.440(0.137) (0.217,0.750)	0.502(0.145) (0.260,0.829)
δ			-0.296(0.069) (-0.437,-0.149)
-2logL	531.5	511.9	499.2
AIC	545.5	527.9	517.2
BIC	567.3	552.8	545.3

Results show that the PSN is the best-fitted while the PN is the second one. The parameter δ differs significantly from 0 based on its confidence interval, and a negative direction of the difference exists. It again supports our claim that the PSN model is more appropriate for the polio data. The Bayesian results differ somewhat for the PSN and Poisson models. Standard deviations for the PSN model are larger, up to 14% and 51%, than those for the PN and Poisson models.

One objective in the analysis of polio data is to investigate whether or not the incidence of polio has been decreasing since 1970. This is indicated by the sign of the regression coefficient β_1. Under the PSN model, the negative sign of the trend term indicates that there is a long term decrease in the number of poliomyelitis cases during the observation period. This finding goes along with results achieved by Davis, Dunsmuir and Wang (2000) and Farrell, MacGibbon and Tomberlin (2007). We also note that the state dependence parameter β_6 is significant in the PSN model, which implies the contribution of the lagged response on prediction, while the PN model does not make such a conclusion.

6.2 Health reform data

The health-care reform data is taken from the German Socio-Economic Panel for the years 1995-1999. The main aim of the study was to investigate whether the number of physician visits by patients decreased after the reform. The data were analysed by Winkelmann (2004), who noted that the number of visits dropped by about 10% on average. Rabe-Hesketh and Skrondal (2012) fitted a PLN regression model on the impact of the 1997 health reform on the number of doctor visits. Then, several studies analysed the data for various purposes (e.g., Van Ophem, 2011). Our data consist of a subset taken from Rabe-Hesketh and Skrondal (2012) and are available in Stata and R software packages. We drop all missing values from the data, giving a subsample of 1,418 women who were employed full time the year before and after the reform.

The response variable is the utilization of health services, as measured by the self-reported number of patient visits to a physician's office three months before the interview. Covariates include an indicator variable for the interview being during the year after the reform versus the year before the reform, centred age in years, person education in years, an indicator for being married, a binary variable for self-reported current health, being classified as 'very poor' or 'poor' (versus 'very good'; 'good' or 'fair'), and the centred logarithm of household income.

The standard Poisson model makes the unrealistic assumption that the number of doctor visits before the reform is independent of the number of visits after the reform for the same person, given the included covariates. A fit of this model gives the ratio of deviance to the degrees-of-freedom equals 3.698, illustrating strong evidence of over-dispersion and suggests fitting alternative models. Thus, we propose fitting mixed Poisson regression models with the following specifications. The counts Y_{it}, conditioned on the effects u_{it} for subject $i = 1, 2, \cdots, 709$ and at time $t = 1, 2$, are taken to be independent $Pois\left(\exp(\theta_{it})\right)$ where

$$\begin{aligned}\theta_{it} = {} & \beta_0 + \beta_1 \text{reform}_{it} + \beta_2 \text{age}_{it} + \beta_3 \text{educ}_{it} + \beta_4 \text{married}_{it} \\ & + \beta_5 \text{badh}_{it} + \beta_6 \text{loginc}_{it} + u_{it}.\end{aligned}$$

We fit PMSN1-PMSN3, PSN, PMN, and PN as competitive models and let $\mathbf{u}_i' = (u_{i1}, u_{i2})$, $\mathbf{V} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and $\mathbf{D} = \sigma_\alpha^2 \mathbf{1}_2 \mathbf{1}_2' + \sigma^2 \mathbf{I}_2$. The heterogeneity effects in PN and PSN models are assumed to be independent. These models are inappropriate. It should come as no surprise since no correlation is allowed for the heterogeneity effects, whereas in reality, it exists. The deviance of PN and PSN models are 5917.7 and 5870.8, respectively. Other findings are dropped here to save space. In addition, the estimate of correlation (11) for model PMSN2 was found to be 0.383 (s.d., 0.068; 95% confidence interval, 0.248–0.517) while for model PMSN3 it was 0.264 (s.d., 0.011; 95% confidence interval, 0.243–0.284). Combining these findings with (11) indicates strong evidence of the correlation between the number of patient visits to a physician's office before and after the reform. The posterior means and standard deviations for the conventional and

proposed models are given in Table 4. In models PMSN1 and PMSN2, the intercept is replaced by β_{0t} for $t = 1,2$.

Table 4: *Posterior summary statistics for proposed PMSN models.*

Model	PMN	PMSN1	PMSN2	PMSN3
	Est(s.d.)	Est(s.d.)	Est(s.d.)	Est(s.d.)
	(95% CI)	(95% CI)	(95% CI)	(95% CI)
β_{01}	1.488(0.122)	0.634(0.062)	0.419(0.061)	0.807(0.062)
	(1.255,1.733)	(0.510,0.757)	(0.303,0.539)	(0.686,0.932)
β_{02}		-0.005(0.004)	1.175(0.103)	
		(-0.013,0.002)	(0.986,1.386)	
β_1	-0.076(0.050)	-0.708(0.076)	-0.715(0.076)	-0.249(0.048)
	(-0.174,0.021)	(-0.220,0.170)	(-0.865,-0.571)	(-0.348,-0.155)
β_2	-0.005(0.004)	-0.003(0.003)	-0.002(0.004)	-0.001(0.003)
	(-0.013,0.002)	(-0.009,0.006)	(-0.010,0.006)	(-0.007,0.007)
β_3	-0.004(0.017)	-0.006(0.016)	-0.006(0.017)	-0.011(0.017)
	(-0.040,0.029)	(-0.040,0.026)	(-0.038,0.028)	(-0.044,0.021)
β_4	0.318(0.063)	0.128(0.079)	0.108(0.073)	0.070(0.076)
	(0.192,0.443)	(-0.029,0.281)	(-0.033,0.252)	(-0.077,0.219)
β_5	1.127(0.101)	1.040(0.101)	1.024(0.097)	1.017(0.102)
	(0.926,1.324)	(0.841,1.236)	(0.834,1.217)	(0.818,1.217)
β_6	0.087(0.104)	0.114(0.097)	0.141(0.099)	0.116(0.097)
	(-0.111,0.293)	(-0.077,0.305)	(-0.046,0.342)	(-0.074,0.305)
δ_1		0.331(0.085)	0.375(0.075)	0.469(0.076)
		(0.166,0.496)	(0.221,0.515)	(0.325,0.632)
δ_2		1.007(0.087)	1.047(0.082)	
		(0.841,1.177)	(0.881,1.202)	
σ^2	0.979(0.066)	0.693(0.079)	0.183(0.057)	0.445(0.059)
	(0.854,1.122)	(0.545,0.845)	(0.086,0.301)	(0.341,0.570)
σ^2_α			0.465(0.062)	0.329(0.077)
			(0.347,0.591)	(0.172,0.479)
ρ	0.519(0.054)	0.667(0.086)		
	(0.406,0.618)	(0.664,0.824)		
-2logL	5645.9	5630.8	5627.4	5635.5
AIC	5660.9	5650.8	5647.4	5653.5
BIC	5697.4	5696.4	5693.0	5694.6

Table 4 also shows Bayes estimates of σ^2 and ρ with their 95% confidence intervals (CI) for models PMSN1 and PMN. That is, with 95% probability σ^2 lies between (0.854,1.122), for example. These facts reveal that much variability exists for the number of visits after the reform. Similarly, all skewness parameters differ significantly from 0 in a positive direction, showing that the distribution of heterogeneity effects is skewed

right. It indicates that models PMSN1, PMSN2, and PMSN3 are more appropriate than the PMN model. Also, according to the deviance and the information criteria values reported in Table *4*, we find that the PMSN2 fits the data better than all competitive models.

Furthermore, in model PMSN2, age, education, married, and loginc are not statistically significant. However, the health care reform is negatively related to the number of visits, meaning that reform makes a decrease in the expected number of visits. Moreover, the badh coefficient is significant and positive, meaning that patients with having bad health make an increase in visits.

7 Concluding remarks

The analysis of correlated counts is challenging since suitable discrete multivariate distributions that can provide appropriate correlation structure are not always available. In longitudinal studies, the problem is addressed by letting the counts be independent Poisson variates conditioned on a vector of correlated heterogeneity effects. The correlation between the count variables is then incorporated in the resulting likelihood functions. In the paper, the correlation was taken into account by adopting that the random effects followed the multivariate skew-normal distribution with various structures for the skewness parameters. The modelling strategy allows for both positive and negative correlations among the subsequent counts. Empirical findings showed that the proposed modelling strategy had many potentials over conventional models. The paper used an accessible technique to compute the AIC and BIC values by plugging in Bayes estimates at the underlying marginal likelihoods. An interesting subject to future work is to use other Bayesian models comparison. Also, an extension of mixed modelling to the multivariate skew-normal random-effects is encouraged for non-Poisson correlated responses when over-dispersion occurs.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88(422), 669-679.

Azzalini, A. (1985). A class of distributions which includes the normal ones. Scandinavian Journal of Statistics 12(2), 171-178.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. Biometrika 83(4),715-726.

Balakrishnan, N. and Peng, Y. (2006). Generalized gamma frailty model. Statistics in Medicine 25(16),2797-2816.

Bulmer, M.G. (1974). On fitting the Poisson log-normal distribution to species-abundance data. Biometrics 30(1),101-110.

Cameron, A.C. and Trivedi, P.K. (2013). Regression Analysis of Count Data. Cambridge University Press, Cambridge.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. The American Statistician 49(4),327-335.

Chib, S. and Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. Journal of Business and Economic Statistics 19(4),428-435.

Davis, R.A., Dunsmuir, W.T. and Wang, Y. (2000). On autocorrelation in a Poisson regression model. Biometrika 87(3),491-505.

Davis, R.A. and Wu, R.A. (2009). A negative binomial model for time series of counts. Biometrika 96(3),735-749.

El-Basyouny, K. and Sayed, T. (2009). Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention 41(4),820-828.

Farrell, P.J., MacGibbon, P. and Tomberlin, T.J. (2007). A hierarchical Bayes approach to estimation and prediction for time series of counts. Brazilian Journal of Probability and Statistics 21(2),187-202.

Fokianos, K. and Fried, R. (2012). Interventions in log-linear Poisson autoregression. Statistical Modelling 12(4),299-322.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Journal of the Royal Statistical Society, Series C(Applied Statistics) 41(2),337-348.

Gonzales-Barron, U. and Butler, F. (2011). A comparison between the discrete Poisson-gamma and Poisson-lognormal distributions to characterise microbial counts in foods. Food Control 22(8),1279-1286.

Guo, J.Q. and Trivedi, P.K. (2002). Flexible parametric models for long-tailed patent count distributions. Oxford Bulletin of Economics and Statistics 64(1),63-82.

Holgate, P. (1970). The modality of some compound Poisson distributions. Biometrika 57(3),666-667.

Izsák, R. (2008). Maximum likelihood fitting of the Poisson lognormal distribution. Environmental and Ecological Statistics 15(2),143-156.

Kang, J. and Lee, S. (2014). Parameter change test for Poisson autoregressive models. Scandinavian Journal of Statistics 41(4),1136-1152.

Karlis, D. and Xekalaki, E. (2005). Mixed Poisson distributions. International Statistical Review 73(1),35-58.

Kuba, M. and Panholzer, A. (2016). On moment sequences and mixed Poisson distributions. Probability Surveys 13,89-155.

Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009). The BUGS project, Evolution, critique and future directions. Statistics in Medicine 28(25),30-49.

Miller, G. (2007). Statistical Modeling of Poisson/Log-Normal Data. Radiation Protection Dosimetry Advance Access Published January 124(2),1-9.

Molenberghs, G., Verbeke, G. and Demétrio, C.G. (2007). An extended random-effects approach to modelling repeated, over-dispersed count data. Lifetime Data Analysis 13(4),513-531.

Montesinos-López, O.A., Montesinos-López, A., Crossa, J., Toledo, F.H., Montesinos-López, J.C., Singh, P., Juliana, p. and Salinas-Ruiz, J. (2017). A Bayesian Poisson-lognormal model for count data for multiple-trait multiple-environment genomic-enabled prediction. G3, Genes, Genomes, Genetics 7(5);1595-1606.

Nadarajah, S. and Kotz, S. (2006a). Compound mixed Poisson distributions I. Scandinavian Actuarial Journal 3,141-162.

Nadarajah, S. and Kotz, S. (2006b). Compound mixed Poisson distributions II. Scandinavian Actuarial Journal 3,163-181.

Oh, M.S. and Lim, Y.B. (2001). Bayesian analysis of time series Poisson data. Journal of Applied Statistics 28(2),259-271.

Rabe-Hesketh, S. and Skrondal, A. (2012). Multilevel and Longitudinal Modeling Using Stata, Volume II: Categorical Responses, Counts, and Survival, 3rd edition, Stata Press, College Station, Texas.

Rizzato, F.B., Leandro, R.A., Demétrio, C.G. and Molenberghs, G. (2016). A Bayesian approach to analyse overdispersed longitudinal count data. Journal of Applied Statistics 43(11),2085-2109.

Sahu, S.K., Dey, D.K. and Branco, M. (2003). A new class of multivariate distributions with applications to Bayesian regression models. The Canadian Journal of Statistics 31(2),129-150.

Van Ophem, H. (2011). The frequency of visiting a doctor, is the decision to go independent of the frequency?. Journal of Applied Econometrics 26(5),872-879.

Williams, M.S. and Ebel, E.D. (2012). Methods for fitting the Poisson-lognormal distribution to microbial testing data. Food Control 27(1),73-80.

Winkelmann, R. (2004). Health care reform and the number of doctor visits an econometric analysis. Journal of Applied Econometrics 19(4),455-472.

Wu, H., Deng, X. and Ramakrishnan, N. (2018). Sparse estimation of multivariate Poisson log-normal models from count data. Statistical Analysis and Data Mining, The ASA Data Science Journal 11(2),66-77.

Zeger, S.L. (1988). A regression model for time series of counts. Biometrika 75(4),621-629.