ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

REVSTAT
Statistical Journal

Special issue on
«AISC 2016: Advances in Interdisciplinary Statistics and Combinatorics»

Guest Editors:
Sat Gupta
Kumer Pial Das

Volume 16, No.2
April 2018

REVSTAT

Statistical Journal

Catalogação Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E., 2003- . - 30 cm
Semestral. - Continuação de : Revista de Estatística = ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726

CREDITS

- **PUBLISHER**
  - *Instituto Nacional de Estatística, I.P. (INE, I.P.)*
    Av. António José de Almeida, 2
    1000-043 LISBOA
    PORTUGAL
    Tel.: + 351 21 842 61 00
    Fax: + 351 21 845 40 84
    Web site: http://www.ine.pt
    Customer Support Service
    + 351 218 440 695
- **COVER DESIGN**
  - *Mário Bouçadas, designed on the stain glass window at INE by the painter Abel Manta*
- **LAYOUT AND GRAPHIC DESIGN**
  - *Carlos Perpétuo*
- **PRINTING**
  - *Instituto Nacional de Estatística, I.P.*
- **EDITION**
  - *140 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*
- **PRICE [VAT included]**
  - *€ 9,00*

PREFACE

This special issue of *REVSTAT — Statistical Journal* features specially invited papers from those who presented at the International Conference on Advances in Interdisciplinary Statistics and Combinatorics held at the University of North Carolina at Greensboro, USA during September 30 – October 2, 2016. The contributions to this special issue cover several very significant areas of statistics such as Bayesian mixture models, non-parametric predictive inference, sampling, and survival analysis.

The guest editors are grateful to the contributors to this issue as well as the editors of *REVSTAT* for their support during the review process. We also wish to acknowledge the help of the referees who reviewed the papers very promptly and diligently.

Guest Editors:

Sat Gupta
Professor
Department of Mathematics and Statistics,
University of North Carolina at Greensboro,
Greensboro, NC, USA

Kumer Pial Das
Professor
Department of Mathematics,
Lamar University,
Beaumont, TX, USA

INDEX

Abstracted/indexed in: *Current Index to Statistics, DOAJ, Google Scholar, Journal Citation Reports/Science Edition, Mathematical Reviews, Science Citation Index Expanded®, SCOPUS* and *Zentralblatt für Mathematic.*

REVSTAT – Statistical Journal
Volume 16, Number 2, April 2018, 167–185

NONPARAMETRIC PREDICTIVE INFERENCE FOR REPRODUCIBILITY OF TWO BASIC TESTS BASED ON ORDER STATISTICS

Authors: Frank P.A. Coolen
– Department of Mathematical Sciences, Durham University, Durham, UK
frank.coolen@durham.ac.uk

Hana N. Alqifari
– Department of Mathematics, Qassim University, Buraidah, Saudi Arabia

Received: February 2017 Revised: June 2017 Accepted: July 2017

Abstract:

- Reproducibility of statistical hypothesis tests is an issue of major importance in applied statistics: if the test were repeated, would the same overall conclusion be reached, that is rejection or non-rejection of the null hypothesis? Nonparametric predictive inference (NPI) provides a natural framework for such inferences, as its explicitly predictive nature fits well with the core problem formulation of a repeat of the test in the future. NPI is a frequentist statistics method using relatively few assumptions, made possible by the use of lower and upper probabilities. For inference on reproducibility of statistical tests, NPI provides lower and upper reproducibility probabilities (RP). In this paper, the NPI-RP method is presented for two basic tests using order statistics, namely a test for a specific value for a population quantile and a precedence test for comparison of data from two populations, as typically used for experiments involving lifetime data if one wishes to conclude before all observations are available.

Key-Words:

- *lower and upper probabilities; nonparametric predictive inference; precedence test; quantile test; reproducibility.*

AMS Subject Classification:

- 60A99, 62G99, 62P30.

1. INTRODUCTION

Testing of hypotheses is one of the main tools in statistics and crucial in many applications. While many different tests have been developed for a wide range of scenarios, the aspect of reproducibility of tests has long been neglected: the question addressed is whether or not a test, if it were repeated under the same circumstances, would lead to the same overall conclusion, that is rejection or non-rejection of the null hypothesis. Recently, this topic has started to gain attention, in particular through the publication of a 'handbook on reproducibility' [4] which provides a collection of papers on the issue. Nevertheless, whilst hypothesis testing is mainly seen as a frequentist statistics procedure, the classic frequentist framework is not suited for inference on reproducibility as this is neither an estimation nor a testing problem. The very nature of reproducibility is predictive, namely given the results of one test one wishes to predict the outcome of a possible future test. Coolen and Bin Himd [11] presented nonparametric predictive inference (NPI) for reproducibility of some basic tests, with more attention to this topic in the PhD thesis of Bin Himd [8], these publications also provide a critical discussion of earlier methods for reproducibility presented in the literature.

This paper contributes to development of NPI for reproducibility by considering two tests based on order statistics, namely a one sample quantile test and a two sample precedence test. Central to these inferences are NPI results for future order statistics [12]. This paper provides a concise presentation of NPI for the quantile and basic precedence test, further details, examples and discussion are included in the PhD thesis of Alqifari [1].

This paper is organized as follows. Section 2 provides a brief introduction to NPI, including key results on NPI for future order statistics as used in this paper. Section 3 discusses aspects of reproducibility of statistical tests and explains the NPI perspective on such inferences. Section 4 presents the NPI approach to reproducibility of a basic quantile test. Section 5 considers a precedence test used for comparison of two populations. Some concluding remarks are given in Section 6. All computations in this paper were performed using the statistical software R.

2. NONPARAMETRIC PREDICTIVE INFERENCE

Nonparametric predictive inference (NPI) [5, 10] is a statistical framework which uses few modelling assumptions, with inferences explicitly in terms of future observations. For real-valued random quantities attention has thus far been

mostly restricted to a single future observation, although multiple future observations have been considered for some NPI methods, e.g. in statistical process control [2, 3].

Assume that we have real-valued ordered data $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, with $n \geq 1$. For ease of notation, define $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$, or at other known lower and upper bounds of the range of possible values for these random quantities. The n observations create a partition of the real-line into $n+1$ intervals $I_j = (x_{(j-1)}, x_{(j)})$ for $j = 1, ..., n+1$. We assume throughout this paper that ties do not occur. If we wish to allow ties, also between past and future observations, we could use closed intervals $[x_{(j-1)}, x_{(j)}]$ instead of these open intervals I_j, the difference is rather minimal and to keep presentation easy we have opted not to do this here. We are interested in $m \geq 1$ future observations, X_{n+i} for $i = 1, ..., m$. We link the data and future observations via Hill's assumption $A_{(n)}$ [17], or, more precisely, via $A_{(n+m-1)}$ (which implies $A_{(n+k)}$ for all $k = 0, 1, ..., m-2$; we will refer to this generically as 'the $A_{(n)}$ assumptions'), which can be considered as a post-data version of a finite exchangeability assumption for $n+m$ random quantities. The $A_{(n)}$ assumptions imply that all possible orderings of the n data observations and the m future observations are equally likely, where the n data observations are not distinguished among each other and neither are the m future observations. Let $S_j = \#\{X_{n+i} \in I_j,\ i = 1, ..., m\}$, then the $A_{(n)}$ assumptions lead to

$$P\left(\bigcap_{j=1}^{n+1} \{S_j = s_j\}\right) = \binom{n+m}{n}^{-1} \tag{2.1}$$

where s_j are non-negative integers with $\sum_{j=1}^{n+1} s_j = m$. Another convenient way to interpret the $A_{(n)}$ assumptions with n data observations and m future observations is to think that n randomly chosen observations out of all $n+m$ real-valued observations are revealed, following which you wish to make inferences about the m unrevealed observations. The $A_{(n)}$ assumptions then imply that one has no information about whether specific values of neighbouring revealed observations make it less or more likely that a future observation falls in between them. For any event involving the m future observations, Equation (2.1) implies that we can count the number of such orderings for which this event holds. Generally in NPI a lower probability for the event of interest is derived by counting all orderings for which this event has to hold, while the corresponding upper probability is derived by counting all orderings for which this event can hold [5, 10].

In NPI, the $A_{(n)}$ assumptions justify the use of resulting inferences directly as predictive probabilities. Using only precise probabilities, such inferences cannot be used for many events of interest, but in NPI we use the fact, in line with De Finetti's Fundamental Theorem of Probability [14], that corresponding optimal bounds can be derived for all events of interest [5]. These bounds are lower and upper probabilities in the theory of imprecise probability [6]. NPI provides

exactly calibrated frequentist inferences [18], and it has strong consistency properties in theory of interval probability [5]. In NPI the n observations are explicitly used through the $A_{(n)}$ assumptions, yet as there is no use of conditioning as in the Bayesian framework, we do not use an explicit notation to indicate this use of the data. The m future observations must be assumed to result from the same sampling method as the n data observations in order to have full exchangeability. NPI is totally based on the $A_{(n)}$ assumptions, which however should be considered with care as they imply e.g. that the specific ordering in which the data appeared is irrelevant, so accepting $A_{(n)}$ implies an exchangeability judgement for the n observations. It is attractive that the appropriateness of this approach can be decided upon after the n observations have become available. NPI is always in line with inferences based on empirical distributions, which is an attractive property when aiming at objectivity [10].

Let $X_{(r)}$, for $r = 1, ..., m$, be the r-th ordered future observation, so $X_{(r)} = X_{n+i}$ for one $i = 1, ..., m$ and $X_{(1)} < X_{(2)} < \cdots < X_{(m)}$. The following probabilities are derived by counting the relevant orderings and use of Equation (2.1). For $j = 1, ..., n+1$ and $r = 1, ..., m$,

$$P\big(X_{(r)} \in I_j\big) = \binom{j+r-2}{j-1}\binom{n-j+1+m-r}{n-j+1}\binom{n+m}{n}^{-1}. \tag{2.2}$$

For this event NPI provides a precise probability, as each of the $\binom{n+m}{n}$ equally likely orderings of n past and m future observations has the r-th ordered future observation in precisely one interval I_j. As Equation (2.2) only specifies the probabilities for the events that $X_{(r)}$ belongs to intervals I_j, it can be considered to provide a partial specification of a probability distribution for $X_{(r)}$, no assumptions are made about the distribution of the probability masses within such intervals I_j.

Analysis of the probability in Equation (2.2) leads to some interesting results, including the logical symmetry $P(X_{(r)} \in I_j) = P(X_{(m+1-r)} \in I_{n+2-j})$. For all r, the probability for $X_{(r)} \in I_j$ is unimodal in j, with the maximum probability assigned to interval I_{j^*} with $\left(\frac{r-1}{m-1}\right)(n+1) \leq j^* \leq \left(\frac{r-1}{m-1}\right)(n+1)+1$. A further interesting property occurs for the special case where the number of future observations is equal to the number of data observations, so $m = n$. In this case, $P(X_{(r)} < x_r) = P(X_{(r)} > x_r) = 0.5$ holds for all $r = 1, ..., m$. This fact can be proven by considering all $\binom{2n}{n}$ equally likely orderings, where clearly in precisely half of these orderings the r-th future observation occurs before the r-th data observation due to the overall exchangeability assumption. The special case $m = n$ plays an important role in this paper as it naturally occurs in analysis of reproducibility of statistical hypothesis tests.

3. REPRODUCIBILITY OF STATISTICAL TESTS

Statistical hypothesis testing is used in many application areas and normally results in either non-rejection of the stated null hypothesis or its rejection in favour of a stated alternative, at a predetermined level of significance. Whilst this procedure is embedded in the successful long-standing tradition of statistics, a related aspect that had received relatively little attention in the literature until recently is the reproducibility of such tests: if the test were repeated, would it lead to the same overall conclusion? Attention to problems with reproducibility, including problems with understanding of concepts by practitioners in application areas, was raised by Goodman [16] and Senn [21]. Methods for addressing reproducibility, proposed in the literature since then, have mainly shown that the classical frequentist framework of statistics may not be immediately suitable for inference on test reproducibility (see [11] for a discussion of such proposed methods). Recently, many aspects of reproducibility, including some attention to statistical methods, have been discussed in a volume dedicated to this topic [4].

The reproducibility probability (RP) for a test is the probability for the event that, if the test is repeated based on an experiment performed in the same way as the original experiment, the test outcome, that is either rejection of the null-hypothesis or not, will be the same. In practice, focus may often be on reproducibility of tests in which the null-hypothesis is rejected, for example because significant effects tend to lead to new treatments in medical applications. However, also if the null-hypothesis is not rejected it is important to have a meaningful assessment of the reproducibility of the test. Note that RP is assessed knowing the outcome of the first, actual experiment, which consists of the actual observations, so not only the value of a sufficient test statistic or even just the conclusion on rejection or non-rejection of the null-hypothesis. This is important as the RP will vary with different experiment outcomes, which is logical and will lead to higher RP if the data supported the original test conclusion more strongly. A sufficient test statistic, if of reduced dimension compared to the full data set, does not provide suitable input for the NPI method, hence the use of the full data set is required for the inferences considered in this paper.

Coolen and Bin Himd [11] introduced NPI for RP, denoted by NPI-RP, by considering some basic nonparametric tests: the sign test, Wilcoxon's signed rank test, and the two sample rank sum test. For these inferences NPI for Bernoulli quantities [9] and for real-valued observations [5] were used. This did not lead to precise valued reproducibility probabilities but to NPI lower and upper reproducibility probabilities, denoted by $\underline{RP}$ and $\overline{RP}$, respectively. For these tests analytic methods were presented to calculate the NPI lower and upper probabilities for test reproducibility. To enable NPI for more complex test scenarios, the NPI-bootstrap method can be used, as introduced and illustrated by Bin Himd [8] for the Kolmogorov–Smirnov test.

This paper presents NPI-RP for two classical tests which are based on order statistics, namely a one sample quantile test (Section 4) and a two sample precedence test (Section 5). For these inferences, NPI for future order statistics [12] is used, as briefly reviewed in Section 2. We assume that the first, actual experiment led to ordered real-valued observations $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. As we consider an imaginary repeat of this experiment, we use NPI for $m = n$ future ordered observations, henceforth denoted by $X^f_{(1)} < X^f_{(2)} < \cdots < X^f_{(n)}$, with the superscript f used to emphasize that we consider future order statistics.

4. QUANTILE TEST

The quantile test is a basic nonparametric test for the value of a population quantile [15]. Let κ_p denote the $100 \times p$-th quantile of an unspecified continuous distribution, for $0 \leq p \leq 1$. On the basis of a sample of observations of independent and identically distributed random quantities X_i, $i = 1, ..., n$, we consider the one-sided test of null-hypothesis H_0: $\kappa_p = \kappa_p^0$ versus alternative H_1: $\kappa_p > \kappa_p^0$, for a specified value κ_p^0. We restrict attention in this paper to NPI for reproducibility of this one-sided quantile test. The corresponding methodology for the two-sided test follows the same steps and is included in the PhD thesis of Alqifari [1], where also some more discussion and examples are given of the tests presented in this paper. Actually, there is an interesting issue about two-sided tests in such scenarios, that requires some further thought. If the original test leads to rejection of the null hypothesis due to a relatively small value of the test statistic, would one consider the test result to be reproduced if a future test leads to rejection due to a relatively large value of the test statistic, so in the other tail of the statistic's distribution under H_0? Technically perhaps this is the case, but on the basis of the combined evidence of the two tests one would probably want to investigate the whole setting further and not regard the second test as confirming the results of the first test. This is left as a topic for consideration.

Under H_0, κ_p^0 is the $100 \times p$-th quantile of the distribution function of the X_i, so $P(X_i \leq \kappa_p^0 \,|\, H_0) = p$. Define the random variable K as the number of X_i in the sample of size n that are less than or equal to κ_p^0, that is

$$K = \sum_{i=1}^{n} \mathbf{1}\{X_i \leq \kappa_p^0\}$$

with $\mathbf{1}\{A\} = 1$ if A is true and $\mathbf{1}\{A\} = 0$ if A is not true. A logical test rule is to reject H_0 if $X_{(r)} > \kappa_p^0$, where $X_{(r)}$ is the r-th ordered observation in the sample (ordered from small to large), for a suitable value of r corresponding to a chosen significance level, so if $K \leq r - 1$. For significance level α, r is the largest integer

such that

$$P\big(X_{(r)} > \kappa_p^0 \,|\, H_0\big) \;=\; \sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i} \;\leq\; \alpha \;.$$

For large sample sizes the Normal distribution approximation to the Binomial distribution can be used in order to determine the appropriate value of r.

For a given data set $x_1, ..., x_n$, the test statistic of the one-sided quantile test as defined above is the number of observations less than or equal to κ_p^0, denoted by

$$k \;=\; \sum_{i=1}^{n} \mathbf{1}\big\{x_i \leq \kappa_p^0\big\} \;.$$

For the value r derived as discussed above, H_0 is rejected if and only if $k \leq r-1$.

Based on such data and the result of the actual hypothesis test, that is whether the null hypothesis is rejected in favour of the alternative hypothesis or not, NPI can be applied to study the reproducibility of the test. First we consider the case where $k \leq r-1$, so the original test leads to rejection of H_0. Reproducibility of this test result is therefore the event that, if the test were repeated, also leading to n observations, then that would also lead to rejection of H_0. Using the notation for future observations introduced in Section 3, this would occur if the r-th ordered observation of the future sample exceeds κ_p^0. The NPI lower and upper reproducibility probabilities for this event, as function of $k \leq r-1$, are

$$\underline{RP}(k) \;=\; \underline{P}\big(X_{(r)}^f > \kappa_p^0 \,|\, k\big) \;=\; \sum_{j=1}^{n+1} \mathbf{1}\big\{x_{j-1} > \kappa_p^0\big\}\, P\big(X_{(r)}^f \in I_j\big)$$

and

$$\overline{RP}(k) \;=\; \overline{P}\big(X_{(r)}^f > \kappa_p^0 \,|\, k\big) \;=\; \sum_{j=1}^{n+1} \mathbf{1}\big\{x_j > \kappa_p^0\big\}\, P\big(X_{(r)}^f \in I_j\big) \;,$$

respectively. Note that the dependence of these lower and upper probabilities on the value k is not explicit in the notation used for the terms on the right-hand side, but is due to the number of data x_j that exceed κ_p^0. It is easily shown that $\underline{P}(X_{(r)}^f > \kappa_p^0 \,|\, k) = \overline{P}(X_{(r)}^f > \kappa_p^0 \,|\, k+1)$, which leads to $\underline{RP}(k) = \overline{RP}(k+1)$ for values of k leading to rejection of H_0.

If the original test does not lead to rejection of H_0, so if $k \geq r$, then reproducibility of the test is the event that the null hypothesis would also not get rejected in the future test. The NPI lower and upper reproducibility probabilities for this event, as function of $k \geq r$, are

$$\underline{RP}(k) \;=\; \underline{P}\big(X_{(r)}^f \leq \kappa_p^0 \,|\, k\big) \;=\; \sum_{j=1}^{n+1} \mathbf{1}\big\{x_j \leq \kappa_p^0\big\}\, P\big(X_{(r)}^f \in I_j\big)$$

and

$$\overline{RP}(k) \;=\; \overline{P}\big(X^f_{(r)} \le \kappa^0_p \,|\, k\big) \;=\; \sum_{j=1}^{n+1} \mathbf{1}\big\{x_{j-1} \le \kappa^0_p\big\}\, P\big(X^f_{(r)} \in I_j\big)\;,$$

respectively. It is easily seen that $\underline{RP}(k) = \overline{RP}(k-1)$ for values of k such that $k-1$ leads to H_0 not being rejected. If an actual observation in the original test is exactly equal to the specified value κ^0_p, then the NPI method would actually provide a precise reproducibility probability. We do not consider this further as the test hypotheses must always be specified without consideration of the actual test data, hence this case is extremely unlikely to occur; for some further discussion see [1].

The minimum value that can occur for the NPI lower reproducibility probabilities for this one-sided quantile test, following either rejection or non-rejection of the null hypothesis in the original test, is equal to 0.5. This follows directly from the formulae for the NPI lower reproducibility probabilities given above, together with $P(X_{(r)} < x_r) = P(X_{(r)} > x_r) = 0.5$ as explained in Section 2. The NPI upper reproducibility probabilities can be equal to one. This occurs when all observations in the original test are less than κ^0_p, so $k = n$, in which case the original test let to H_0 not being rejected for all values of r (so for all order statistics considered, hence for any level of significance); this reflects that, with no evidence in the original data in favour of the possibility that the data values can actually exceed κ^0_p, one cannot exclude the possibility that no future observations could exceed this value. Note that the corresponding NPI lower reproducibility probability will be less than one, reflecting that the original data set only provides limited information, this lower probability will increase towards one as function of n. The upper reproducibility probability is also equal to one if all observations in the original test are greater than κ^0_p, so $k = 0$, for which case the reasoning is similar to that above but of course now with H_0 being rejected.

Example 1. Suppose that the original test has sample size $n = 15$ and we are interested in testing the null hypothesis that the third quartile, so the 75% quantile, of the underlying distribution is equal to a specified value $\kappa^0_{0.75}$ against the alternative hypothesis that this third quartile is greater than $\kappa^0_{0.75}$, tested at significance level $\alpha = 0.05$. Using the Binomial distribution for the classical quantile test, this leads to the rule that H_0 is rejected if $x_{(8)} > \kappa^0_{0.75}$ and H_0 is not rejected if $x_{(8)} < \kappa^0_{0.75}$. Note that we do not discuss the case $x_{(8)} = \kappa^0_{0.75}$ which is slightly different as the NPI approach leads to precise probabilities instead of lower and upper probabilities (see [1]), it is also of little practical relevance.

Table 1 presents the NPI lower and upper reproducibility probabilities for all values of k, which is the number of observations in the original test which are less than $\kappa^0_{0.75}$. If $k \le 7$ then the original test leads to H_0 being rejected while it is not rejected for $k \ge 8$. Hence, the NPI lower and upper reproducibility probabilities are for the events $X^f_{(8)} > \kappa^0_{0.75}$ and $X^f_{(8)} < \kappa^0_{0.75}$, respectively. This

table illustrates the logical fact that the worst reproducibility is achieved for k at the threshold values 7 and 8, with increasing RP values when moving away from these values, leading to maximum NPI-RP values for $k = 0$ and $k = 15$. Because for this test the threshold between rejecting and not rejecting H_0 is between $k = 7$ and $k = 8$ out of $n = 15$ observations, the NPI-RP values are symmetric, that is the same for $k = j$ and $k = 15 - j$ for $j = 0, 1, ..., 7$ in Table 1.

Table 1: NPI-RP for third quartile, $n = 15$ and $\alpha = 0.05$.

k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$
0	0.9989	1	6	0.6424	0.7689	12	0.9359	0.9749
1	0.9929	0.9989	7	0.5	0.6424	13	0.9749	0.9929
2	0.9749	0.9929	8	0.5	0.6424	14	0.9929	0.9989
3	0.9359	0.9749	9	0.6424	0.7689	15	0.9989	1
4	0.8682	0.9359	10	0.7689	0.8682			
5	0.7689	0.8682	11	0.8682	0.9359			

Table 2 presents NPI-RP values for the quantile test considering the median, so the 50% quantile, again with sample size $n = 15$ and testing the null hypothesis that the median is equal to a specified value $\kappa_{0.5}^0$ against the one-sided hypothesis that it is greater than $\kappa_{0.5}^0$, at level of significance $\alpha = 0.05$. This leads to the test rule that H_0 is rejected if the number k of observations that are smaller than $\kappa_{0.5}^0$ is less than or equal to 3, and H_0 is not rejected if $k \geq 4$. Note that throughout this paper, precise values 0.5 and 1 are presented without additional decimals, so the values 1.0000 are actually less than 1 but rounded upwards. Of course, these NPI-RP values are not symmetric, and reproducibility becomes very likely for initial test results with a substantial number of observations less than $\kappa_{0.5}^0$. But rejection of H_0, which occurs for $k \leq 3$ and is often of main practical relevance, has relatively low NPI-RP values.

Table 2: NPI-RP for median, $n = 15$ and $\alpha = 0.05$.

k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$
0	0.9502	1	6	0.7865	0.8775	12	0.9986	0.9997
1	0.8352	0.9502	7	0.8775	0.9359	13	0.9997	0.9999
2	0.6743	0.8352	8	0.9359	0.9698	14	0.9999	1.0000
3	0.5	0.6743	9	0.9698	0.9873	15	1.0000	1
4	0.5	0.6592	10	0.9873	0.9954			
5	0.6592	0.7865	11	0.9954	0.9986			

Tables 3 and 4 present the NPI-RP results for the same one-sided quantile test on the third quartile for $n = 30$, at significance levels $\alpha = 0.05$ and $\alpha =$

0.01, respectively. Using the Normal distribution approximation, the test rule for $\alpha = 0.05$ is to reject H_0 that this third quartile is equal to $\kappa^0_{0.75}$ in favour of the alternative hypothesis that it is greater than $\kappa^0_{0.75}$ if $k \leq 18$ and not to reject it if $k \geq 19$, where k is again the number of observations less than $\kappa^0_{0.75}$. For $\alpha = 0.01$, H_0 is rejected if $k \leq 16$ and not rejected if $k \geq 17$. The change in level of significance α leads obviously to change of the rejection threshold, with H_0 being rejected for a smaller range of values k in case of smaller value of α. Comparison of Table 3 with Table 1 shows that the larger sample size tends to lead to slightly less imprecision, that is the difference between corresponding upper and lower probabilities, this is e.g. shown by considering the upper probabilities $\overline{RP}(k)$ for the values of k next to the rejection thresholds, so corresponding to $\underline{RP}(k) = 0.5$.

Table 3: NPI-RP for third quartile, $n = 30$ and $\alpha = 0.05$.

k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$
0	1.0000	1	11	0.9651	0.9811	22	0.7941	0.8666
1	1.0000	1.0000	12	0.9398	0.9651	23	0.8666	0.9210
2	1.0000	1.0000	13	0.9023	0.9398	24	0.9210	0.9580
3	1.0000	1.0000	14	0.8503	0.9023	25	0.9580	0.9805
4	0.9999	1.0000	15	0.7826	0.8503	26	0.9805	0.9923
5	0.9998	0.9999	16	0.6995	0.7826	27	0.9923	0.9976
6	0.9993	0.9998	17	0.6038	0.6995	28	0.9976	0.9995
7	0.9981	0.9993	18	0.5	0.6038	29	0.9995	0.9999
8	0.9956	0.9981	19	0.5	0.6054	30	0.9999	1
9	0.9905	0.9956	20	0.6054	0.7056			
10	0.9811	0.9905	21	0.7056	0.7941			

Table 4: NPI-RP for third quartile, $n = 30$ and $\alpha = 0.01$.

k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$	k	$\underline{RP}(k)$	$\overline{RP}(k)$
0	1.0000	1	11	0.9023	0.9406	22	0.9101	0.9483
1	1.0000	1.0000	12	0.8493	0.9023	23	0.9483	0.9731
2	1.0000	1.0000	13	0.7805	0.8493	24	0.9731	0.9875
3	0.9999	1.0000	14	0.6971	0.7805	25	0.9875	0.9949
4	0.9995	0.9999	15	0.6019	0.6971	26	0.9949	0.9983
5	0.9986	0.9995	16	0.5	0.6019	27	0.9983	0.9995
6	0.9964	0.9986	17	0.5	0.6026	28	0.9995	0.9999
7	0.9916	0.9964	18	0.6026	0.6995	29	0.9999	1.0000
8	0.9824	0.9916	19	0.6995	0.7852	30	1.0000	1
9	0.9664	0.9824	20	0.7852	0.8559			
10	0.9406	0.9664	21	0.8559	0.9101			

5. PRECEDENCE TEST

As a second example of NPI for reproducibility of a statistical test based on order statistics we consider a basic nonparametric precedence test. Such a test, first proposed by Nelson [19], is typically used for comparison of two groups of lifetime data, where one wishes to reach a conclusion before all units on test have failed. The test is based on the order of the observed failure times for the two groups, and typically leads to, possibly many, right-censored observations at the time when the test is ended. Balakrishnan and Ng [7] present a detailed introduction and overview of precedence testing, including more sophisticated tests than the basic one considered in this paper. NPI for precedence testing was presented by Coolen-Schrijner *et al.* [13], without consideration of reproducibility. It should be emphasized that we consider here the NPI approach for reproducibility of a classical precedence test, so not of the NPI approach to precedence testing [13].

We consider the classical scenario with two independent samples. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n_x)}$ be the ordered real-valued observations in a sample of size n_x drawn randomly from a continuously distributed population, which we refer to as the X population, with a probability distribution depending on location parameter λ_x. Similarly, let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n_y)}$ be the ordered real-valued observations in a sample of size n_y drawn randomly from another continuously distributed population (the Y population) with a probability distribution which is identical to that of the X population except for its location parameter λ_y. The hypothesis test for the locations of these two populations considered here is $H_0\colon \lambda_x = \lambda_y$ versus $H_1\colon \lambda_x < \lambda_y$, which is to be interpreted such that, under H_1, observations from the Y population tend to be larger than observations from the X population.

The precedence test considered in this paper, for this specific hypothesis test scenario, is as follows. Given n_x and n_y, one specifies the value of r, such that the test is ended at, or before, the r-th observation of the Y population. For specific level of significance α, one determines the value k (which therefore is a function of α and of r) such that H_0 is rejected if and only if $X_{(k)} < Y_{(r)}$. The critical value for k is the smallest integer which satisfies

$$P\big(X_k < Y_r | H_0\big) \;=\; \binom{n_x + n_y}{n_x}^{-1} \sum_{j=0}^{r-1} \binom{j+k-1}{j} \binom{n_y - j + n_x - k}{n_y - j} \;\leq\; \alpha\,.$$

Note that the test is typically ended at the time $T = \min(X_{(k)}, Y_{(r)})$, with the conclusion that H_0 is rejected in favour of the one-sided alternative hypothesis H_1 specified above if $T = X_{(k)}$ and H_0 is not rejected if $T = Y_{(r)}$. It is of interest to emphasize this censoring; continuing with the original test would make no difference at all to the test conclusion, but further observations would make a difference for the NPI reproducibility results, as will be discussed later.

The NPI approach for reproducibility of this two-sample precedence test considers again the same test scenario applied to future order statistics, and derives the lower and upper probabilities for the event that the same overall test conclusion will be derived, given the data from the original test. This involves the earlier described NPI approach for inference on the r-th future order statistic $Y^f_{(r)}$ out of n_y future observations based on the data from the Y population, and similarly for the k-th future order statistics $X^f_{(k)}$ out of the n_x future observations based on the data from the X population, where the values of r and k are the same as used for the original test (as we assume also the same significance level for the future test). Note, however, that there is a complication: for full specification of the NPI probabilities for these future order statistics, we require the full data from the original test to be available. But, as mentioned, the data resulting from the original precedence test typically has right-censored observations for at least one, but most likely both populations, and these are all just known to exceed the time T at which the original test had ended.

Before we proceed, we discuss this situation in more detail as it is important for the general idea of studying reproducibility of tests. We should emphasize that we have not come across this issue before in the literature, but it seems to be important and more details are provided by Alqifari [1]. There are two perspectives on the study of reproducibility of such precedence tests. First, one can study the test outcome assuming that actually complete data were available, so all n_x and n_y observations of the X and Y populations, respectively, in the original test are assumed to be available. Secondly, one can consider inference for the realistic scenario with the actual data from the original test, so including right-censored observations at time T. The first scenario is the most straightforward for the development of NPI-RP, and we start with this scenario. Then we explain how this first scenario, without additional assumptions, leads to NPI-RP for the second scenario.

The starting point for NPI-RP for the precedence test is to apply NPI for n_x future observations, based on the n_x original test observations from the X population, which are assumed to be fully available, and similarly for n_y future observations based on the n_y observations from the Y population. Using the results presented in Section 2, with notation adapted to indicate the specific populations, the following NPI lower and upper reproducibility probabilities are derived. First, if H_0 is rejected in the original test, so $x_{(k)} < y_{(r)}$, then

$$\underline{RP} = \underline{P}\big(X^f_{(k)} < Y^f_{(r)}\big) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\big\{x_{(j_x)} < y_{(j_y-1)}\big\}\, P\big(X^f_{(k)} \in I^x_{j_x}\big)\, P\big(Y^f_{(r)} \in I^y_{j_y}\big)\,,$$

$$\overline{RP} = \overline{P}\big(X^f_{(k)} < Y^f_{(r)}\big) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\big\{x_{(j_x-1)} < y_{(j_y)}\big\}\, P\big(X^f_{(k)} \in I^x_{j_x}\big)\, P\big(Y^f_{(r)} \in I^y_{j_y}\big)\,.$$

If H_0 is not rejected in the original test, so $x_{(k)} > y_{(r)}$, then

$$\underline{RP} = \underline{P}\big(X^f_{(k)} > Y^f_{(r)}\big) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\big\{x_{(j_x-1)} < y_{(j_y)}\big\}\, P\big(X^f_{(k)} \in I^x_{j_x}\big)\, P\big(Y^f_{(r)} \in I^y_{j_y}\big)\,,$$

$$\overline{RP} = \overline{P}\big(X^f_{(k)} > Y^f_{(r)}\big) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\big\{x_{(j_x)} < y_{(j_y-1)}\big\}\, P\big(X^f_{(k)} \in I^x_{j_x}\big)\, P\big(Y^f_{(r)} \in I^y_{j_y}\big)\,.$$

The following general results for this NPI lower and upper reproducibility probabilities are easily derived [1]. Both in case of rejecting and not rejecting H_0, the maximum possible value of the NPI upper reproducibility probability is 1. If H_0 was rejected this occurs if $x_{(n_x)} < y_{(1)}$, while if H_0 was not rejected this occurs if $x_{(1)} > y_{(n_y)}$, so both cases lead to maximum reproducibility if the original test data were entirely separated in the sense that either all observations from X population occurred before all observations from the Y population, or the other way around.

In both cases of rejecting or not rejecting H_0 in the original test, the minimum value of the NPI lower reproducibility probability is 0.25. If H_0 was rejected, this occurs if $y_{(r-1)} < x_{(1)}$ and $x_{(k)} < y_{(r)}$ and $y_{(n_y)} < x_{(k+1)}$. If H_0 was not rejected, this occurs if $x_{(k-1)} < y_{(1)}$ and $y_{(r)} < x_{(k)}$ and $x_{(n_x)} < y_{(r+1)}$. Both these smallest possible values for $\underline{RP}$ result from data orderings that, whilst leading to a test conclusion, are least supportive for it, together with the fact that $P(X^f_{(k)} < x_{(k)}) = P(X^f_{(k)} > x_{(k)}) = 0.5$ (and similar for $Y^f_{(r)}$) as discussed in Section 2.

The effect of local changes to the combined ordering of the data of the two populations in the original test is important. Suppose that, for given data for the X and Y populations for the original test, observations $y_{(u)}$ and $x_{(v)}$ are such that $y_{(u)} < x_{(v)}$ and in the combined ordering of all $n_x + n_y$ data they are consecutive. Now suppose that we change these observations, and denote them by $\tilde{y}_{(u)}$ and $\tilde{x}_{(v)}$, respectively, such that they keep their order in the data from their own population but between them change their order, so $\tilde{x}_{(v)} < \tilde{y}_{(u)}$. Then this local change to the combined ordering of the data leads to increase of both the NPI lower and upper probabilities for the event $X_{(k)} < Y_{(r)}$, that is

$$\underline{P}\big(X_{(k)} < Y_{(r)} \,|\, y_{(u)} < x_{(v)}\big) \;<\; \underline{P}\big(X_{(k)} < Y_{(r)} \,|\, \tilde{x}_{(v)} < \tilde{y}_{(u)}\big)\,,$$

$$\overline{P}\big(X_{(k)} < Y_{(r)} \,|\, y_{(u)} < x_{(v)}\big) \;<\; \overline{P}\big(X_{(k)} < Y_{(r)} \,|\, \tilde{x}_{(v)} < \tilde{y}_{(u)}\big)\,.$$

This implies that the NPI-RP inferences for the precedence test depend monotonically on the combined ordering of the original test data, which is an important property to derive such inference for actual tests including right-censored observations, as discussed after the next example.

Example 2. Nelson [20] presents data consisting of six groups of times (in minutes) to breakdown of an insulating fluid subjected to different levels of voltage. To illustrate NPI-RP for the basic precedence test as discussed above, we assume that sample 3 provides data from the X population and sample 6 from the Y population, these times are presented in Table 5. Both samples are of size 10, and we assume that the precedence testing scenario discussed in this section is followed, so we assume that the population distributions may only differ in location parameters, with $H_0\colon \lambda_x = \lambda_y$ tested versus $H_1\colon \lambda_x < \lambda_y$. We assume that $r = 6$, so the test is set up to end at the observation of the sixth failure time for the Y population. We discuss both significance levels $\alpha = 0.05$ and $\alpha = 0.1$. The missing values in Table 5 are only known to exceed 3.83.

Table 5: Times to insulating fluid breakdown.

X sample	0.94	0.64	0.82	0.93	1.08	1.99	2.06	2.15	2.57	*
Y sample	1.34	1.49	1.56	2.10	2.12	3.83	*	*	*	*

For significance level $\alpha = 0.05$, the critical value is $k = 10$, while for $\alpha = 0.1$ this is $k = 9$. Therefore, the provided data will lead, in this precedence test, to rejection of H_0 at 10% level of significance but not to rejection of H_0 at 5% level of significance. For both scenarios, the NPI lower and upper reproducibility probabilities are presented in Table 6, for all of the possible orderings of the right-censored observations. Note that in total 15 observations are available, with 1 value of the X sample and 4 values of the Y sample only known to exceed 3.83.

Table 6: NPI-RP for precedence test on insulating fluid breakdown data.

rank of x_{10}	$\alpha = 0.05$		$\alpha = 0.1$	
	$\underline{RP}$	$\overline{RP}$	$\underline{RP}$	$\overline{RP}$
16	0.3871	0.7814	0.3885	0.7079
17	0.4746	0.8209	0.3490	0.6665
18	0.5496	0.8484	0.3215	0.6309
19	0.6019	0.8627	0.3072	0.6062
20	0.6290	0.8669	0.3029	0.5934

In this table, we give the rank, from the combined ordering of all 20 observations, of the right-censored observation $x_{(10)}$, for example when this is 17 it implies that $y_{(7)} < x_{(10)} < y_{(8)}$. Table 6 presents both the results for $\alpha = 0.05$, in which case H_0 was not rejected in the original test, hence reproducibility is achieved if H_0 is also not rejected in the future test, and the results for $\alpha = 0.05$, in which case

H_0 was rejected so reproducibility also implies rejection of H_0 in the future test. Note that for $\alpha = 0.1$ we still assume that $y_{(6)} = 3.83$ was actually observed, even though the test could have been concluded at time $x_{(9)} = 2.57$ because $x_{(9)} < y_{(6)}$ was conclusive for the test in this case. Table 6 shows that the NPI-RP values are increasing in the combined rank of $x_{(10)}$ for $\alpha = 0.05$ and decreasing for $\alpha = 0.1$, which illustrates the monotonicity of these inferences with regard to changes in ranks of the data as discussed above, as increasing combined rank of $x_{(10)}$ provides more evidence in support of H_0, hence in favour of reproducing the original test result for $\alpha = 0.05$ but against doing so for $\alpha = 0.1$. We notice that the actual rank that $x_{(10)}$ would have among the 20 combined observations has substantial influence on the NPI-RP values. In this example, the imprecision $\overline{RP} - \underline{RP}$ is large. This is due to the relatively small data sets and the fact that two groups of data are compared, with imprecision for the predictive inferences for both groups through the $A_{(n)}$ assumptions for each group.

Thus far, we have studied reproducibility of the basic precedence test from the perspective of having the complete data available, in Example 2 this was illustrated by considering all possible orderings for the right-censored data in the two samples. However, a more realistic perspective is to only use the actual test outcome, without any assumptions on the ordering of the right-censored observations. Using lower and upper probabilities, this can be easily achieved by defining $\underline{RP}$ as the minimum of all NPI lower probabilities for reproducibility over all possible orderings for the right-censored observations, and similarly by defining $\overline{RP}$ as the maximum of all NPI upper probabilities for reproducibility over all possible orderings for the right-censored observations. Hence, in Example 2, this leads to $\underline{RP} = 0.3871$ and $\overline{RP} = 0.8669$ for $\alpha = 0.05$, and $\underline{RP} = 0.3029$ and $\overline{RP} = 0.7079$ for $\alpha = 0.1$. Of course, this leads to increased imprecision compared to every possible specific ordering of the right-censored observations, but it is convenient as no further assumptions about those right-censored observations are required. Furthermore, to derive the NPI-RP values for this perspective one does not need to calculate the corresponding values for each possible combined ordering of right-censored observations, due to the above discussed monotonicity of these inferences. Hence, we always know for which specific ordering of right-censored observations these NPI-RP values are obtained, that is either with all right-censored observations from the X sample occurring before all right-censored observations from the Y sample, or the other way around, depending on the actual outcome of the original test. This perspective is illustrated further in Example 3.

Example 3. We consider again NPI-RP for the precedence test as presented in this section, so with one-sided alternative hypothesis H_1: $\lambda_x < \lambda_y$. Suppose that $n_x = 10$ units of the X population and $n_y = 8$ units of the Y population are put on a life test, where one wants at most two Y units to actually fail, so the value $r = 2$ is chosen. Testing at significance level $\alpha = 0.05$, the critical value is $k = 7$, so H_0 is rejected if $x_{(7)} < y_{(2)}$ while H_0 is not rejected if $y_{(2)} < x_{(7)}$.

Note that, with the test ending at time $\min(x_{(7)}, y_{(2)})$, there are least 3 right-censored X observations and at least 6 right-censored Y observations; this leads to large imprecision in the NPI-RP values. Table 7 presents the NPI lower and upper reproducibility probabilities for this test, for all possible data in the original test, which are indicated through the rankings of all observations until the test is ended, in the combined ranking of the X and Y samples. As indicated, the columns to the left relate to the cases where H_0 is not rejected while the columns to the right relate to the cases where H_0 is rejected. All these NPI-RP values are calculated using the monotonicity with regard to the combined ranks of the right-censored observations, as explained above. These results illustrate the earlier discussed maximum value 1 for $\overline{RP}$ and minimum value 0.25 for $\underline{RP}$. It is particularly noticeable that the NPI lower reproducibility probabilities for this test tend to be small, which is not really surprising due to the large number of right-censored observations resulting from the choice $r = 2$.

Table 7: NPI-RP for precedence test with $n_x = 10$, $n_y = 8$, $r = 2$, $k = 7$ and $\alpha = 0.05$.

H_0 not rejected				H_0 rejected			
X ranks	Y ranks	$\underline{RP}$	$\overline{RP}$	X ranks	Y ranks	$\underline{RP}$	$\overline{RP}$
—	1,2	0.4992	1	1–7	—	0.3833	1
1	2,3	0.4951	0.9988	1–6,8	7	0.3367	0.8833
2	1,3	0.4970	0.9992	1–5,7,8	6	0.2993	0.8425
1,2	3,4	0.4826	0.9924	1–4,6–8	5	0.2739	0.8098
1,3	2,4	0.4884	0.9946	1–3,5–8	4	0.2593	0.7875
2,3	1,4	0.4903	0.9951	1,2,4–8	3	0.2526	0.7748
1–3	4,5	0.4553	0.9733	1,3–8	2	0.2504	0.7690
1–4	5,6	0.4075	0.9314	2–8	1	0.25	0.7670
1–5	6,7	0.3375	0.8582				
1–6	7,8	0.25	0.7509				
2–7	1,8	0.3663	0.8375				

6. CONCLUDING REMARKS

The NPI approach to reproducibility of tests provides many research challenges. It can be developed for many statistical tests, while for some data types (e.g. multivariate data) first NPI requires to be developed further. The test scenarios studied for particular tests may require careful attention, as illustrated by the different perspectives discussed for the precedence test in Section 5. As mentioned, the precedence test scenario discussed in this paper is very basic. Balakrishnan and Ng [7] present a detailed introduction and overview of precedence testing, including more sophisticated tests than the basic one considered

in this paper. In practice, it is important for such tests, and also in general, to also consider the power of the test; thus far this has not yet been considered in the NPI approach for reproducibility of testing. With further development of this approach, we are aiming at guidance on selection of test methods which, for specified level of significance, have good power and good reproducibility properties. This may often require more test data than needed following traditional guidance, but the assurance of good reproducibility is important for many applications and may lead to savings in the longer run by reducing processes, such as development of new medication, to continue on the basis of false test results which may later turn out not to be reproduced in repeated tests under similar circumstances. Further details, examples and discussion of the tests presented in this paper are given in the PhD thesis of Alqifari [1].

ACKNOWLEDGMENTS

This work was carried out while Hana Alqifari was studying for PhD at the Department of Mathematical Sciences, Durham University, supported by the Ministry of Higher Education in Saudi Arabia and Qassim University. The authors gratefully acknowledge detailed comments by three anonymous reviewers which led to improved presentation of the paper. The authors also thank Professor Sat Gupta for the kind invitation to present this work at the International Conference on Advances in Interdisciplinary Statistics and Combinatorics at The University of North Carolina at Greensboro (October 2016), and the subsequent invitation to submit this paper for the conference special issue of this journal.

REFERENCES

[1] Alqifari, H.N. (2017). *Nonparametric Predictive Inference for Future Order Statistics*, PhD Thesis, Durham University (available from `www.npi-statistics.com`).

[2] Arts, G.R.J. and Coolen, F.P.A. (2008). Two nonparametric predictive control charts, *Journal of Statistical Theory and Practice*, **2**, 499–512.

[3] Arts, G.R.J.; Coolen, F.P.A. and van der Laan, P. (2004). Nonparametric predictive inference in statistical process control, *Quality Technology and Quantitative Management*, **1**, 201–216.

[4] Atmanspacher, H. and Maasen, S. (Eds.) (2016). *Reproducibility: Principles, Problems, Practices and Prospects*, Wiley, Hoboken, New Jersey.

[5] Augustin, T. and Coolen, F.P.A. (2004). Nonparametric predictive inference and interval probability, *Journal of Statistical Planning and Inference*, **124**, 251–272.

[6] Augustin, T.; Coolen, F.P.A.; de Cooman, G. and Troffaes, M.C.M. (Eds.) (2014). *Introduction to Imprecise Probabilities*, Wiley, Chichester.

[7] Balakrishnan, N. and Ng, H.K.T. (2006). *Precedence-Type Tests and Applications*, Wiley, Hoboken, New Jersey.

[8] Bin Himd, S. (2014). *Nonparametric Predictive Methods for Bootstrap and Test Reproducibility*, PhD Thesis, Durham University (available from `www.npi-statistics.com`).

[9] Coolen, F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem, *Statistics and Probability Letters*, **36**, 349–357.

[10] Coolen, F.P.A. (2006). On nonparametric predictive inference and objective Bayesianism, *Journal of Logic, Language and Information*, **15**, 21–47.

[11] Coolen, F.P.A. and Bin Himd, S. (2014). Nonparametric predictive inference for reproducibility of basic nonparametric tests, *Journal of Statistical Theory and Practice*, **8**, 591–618.

[12] Coolen, F.P.A.; Coolen-Maturi, T. and Alqifari, H.N. (2018). Nonparametric predictive inference for future order statistics, *Communications in Statistics — Theory and Methods*, **47**, 2527–2548.

[13] Coolen-Schrijner, P.; Maturi, T.A. and Coolen, F.P.A. (2009). Nonparametric predictive precedence testing for two groups, *Journal of Statistical Theory and Practice*, **3**, 273–287.

[14] De Finetti, B. (1974). *Theory of Probability* (2 volumes), Wiley, London.

[15] Gibbons, J.D. and Chakraborti, S. (2010). *Nonparametric Statistical Inference* (5th ed.), Chapman & Hall, Boca Raton, FL.

[16] Goodman, S.N. (1992). A comment on replication, p-values and evidence, *Statistics in Medicine*, **11**, 875–879.

[17] Hill, B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population, *Journal of the American Statistical Association*, **63**, 677–691.

[18] Lawless, J.F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions, *Biometrika*, **92**, 529–542.

[19] Nelson, L.S. (1963). Tables for a precedence life test, *Technometrics*, **124**, 491–499.

[20] Nelson, W.B. (1982). *Applied Life Data Analysis*, Wiley, Hoboken, New Jersey.

[21] Senn, S. (2002). Comment on 'A comment on replication, p-values and evidence', by S.N. Goodman (Letter to the editor), *Statistics in Medicine*, **21**, 2437–2444. With author's reply, pp. 245–247.

REVSTAT – Statistical Journal
Volume 16, Number 2, April 2018, 187–212

MODIFIED SYSTEMATIC SAMPLING WITH MULTIPLE RANDOM STARTS

Authors: Sat Gupta
– Department of Mathematics and Statistics, University of North Carolina, Greensboro, USA
sngupta@uncg.edu

Zaheen Khan
– Department of Mathematics and Statistics, Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan
zkurdu@gmail.com

Javid Shabbir
– Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan
javidshabbir@gmail.com

Received: February 2017 Revised: July 2017 Accepted: August 2017

Abstract:

- Systematic sampling has been facing two problems since its beginning; situational complications, e.g., population size N not being a multiple of the sample size n, and unavailability of unbiased estimators of population variance for all possible combinations of N and n. These problems demand a sampling design that may solve the said problems in a practicable way. In this paper, therefore, a new sampling design is introduced and named as, "Modified Systematic Sampling with Multiple Random Starts". Linear systematic sampling and simple random sampling are the two extreme cases of the proposed design. The proposed design is analyzed in detail and various expressions have been derived. It is found that the expressions for linear systematic sampling and simple random sampling may be extracted from these expressions. Finally, a detailed efficiency comparison is also carried out in this paper.

Key-Words:

- *Modified Systematic Sampling; Linear Systematic Sampling; Simple Random Sampling; Circular Systematic Sampling; Modified Systematic Sampling: Linear Trend.*

AMS Subject Classification:

- 62DO5.

1. INTRODUCTION

Systematic sampling is generally more efficient than Simple Random Sampling (SRS) because SRS may include bulk of units from high density or low density parts of the region, whereas the systematic sampling ensures even coverage of the entire region for all units. In many situations, systematic sampling is also more precise than stratified random sampling. Due to this, researchers and field workers are often inclined towards systematic sampling.

On the other hand, in Linear Systematic Sampling (LSS), we may obtain sample sizes that vary when the population size N is not a multiple of the sample size n, i.e., $N \neq nk$, where k is the sampling interval. However, this problem can be dealt by Circular Systematic Sampling (CSS), Modified Systematic Sampling (MSS) proposed by Khan *et al.* (2013), Remainder Linear Systematic Sampling (RLSS) proposed by Chang and Huang (2000) and Generalized Modified Linear Systematic Sampling Scheme (GMLSS) proposed by Subramani and Gupta (2014). Another well-known and long-standing problem in systematic sampling design is an absence of a design based variance estimator that is theoretically justified and generally applicable. The main reason behind this problem lies in the second-order inclusion probabilities which are not positive for all pairs of units under systematic sampling scheme. It is also obvious that population variance can be unbiasedly estimated if and only if the second-order inclusion probabilities are positive for all pairs of units. To overcome this problem, several alternatives have been proposed by different researchers. However, the simplest one is the use of multiple random starts in systematic sampling. This procedure was adopted by Gautschi (1957) in case of LSS. Later on, Sampath (2009) has considered LSS with two random starts and developed an unbiased estimator for finite-population variance. Sampath and Ammani (2012) further studied the other versions (balanced and centered systematic sampling schemes) of LSS for estimating the finite-population variance. They also discussed the question of determination of the number of random starts. Besides these attempts, the other approaches proposed by different researchers in the past are not much beneficial due to the considerable loss of simplicity.

From the attempts of Gautschi (1957), Sampath (2009), Sampath and Ammani (2012) and Naidoo *et al.* (2016), unbiased estimation of population variance becomes possible just for the case in which $N = nk$. Therefore, to avoid the difficulty in estimation of population variance for the case $N \neq nk$, practitioners are unwillingly inclined towards SRS instead of systematic sampling. Such limitations demand a more generalized sampling design which can play wide-ranging role in the theory of systematic sampling. Thus, in this paper we propose Modified Systematic Sampling with Multiple Random Starts (MSSM). The MSSM ensures unbiased estimation of population variance for the situation where $N \neq nk$.

As one can see, MSS proposed by Khan *et al.* (2013) nicely arranges the population units into k_1 systematic groups each containing s number of units. In MSS, initially a group is selected at random and other $(m-1)$ groups are systematically selected. In this way, a sample of size n consisting of m groups of size s is achieved. Whereas in MSSM, we propose to select all m systematic groups at random to get a sample of size n. Such selection enables us to derive the unbiased variance estimator in systematic sampling. It is interesting to note that LSS and SRS become the extreme cases of MSSM. The MSSM becomes LSS in a situation when N itself is the least common multiple (lcm) of N and n or equivalently $N = nk$, and becomes SRS if lcm is the product of N and n. Because in the case when $N = nk$ we are selecting $m = 1$ group at random which resembles LSS. Whereas, if lcm is the product of N and n we have N groups each containing only one unit from which we are selecting n groups at random in MSSM, which is similar to SRS. In case of LSS, variance estimation can be easily dealt with by Gautschi (1957), Sampath (2009), Sampath and Ammani (2012) and Naidoo *et al.* (2016); whereas the worst case of MSSM is SRS, where unbiased variance estimation can be done using SRS approach.

2. MODIFIED SYSTEMATIC SAMPLING WITH MULTIPLE RANDOM STARTS

Suppose, we have a population of size N, the units of which are denoted by $\{U_1, U_2, U_3, ..., U_N\}$. To select a sample of size n from this population, we will arrange N units into $k_1 = L/n$ (where L is the least common multiple of N and n) groups, each containing $s = N/k_1$ elements. The partitioning of groups is shown in Table 1. A set of $m = L/N$ groups from these k_1 groups are selected using simple random sampling without replacement to get a sample of size $ms = n$.

Table 1: Labels of population units arranged in MSSM.

		Labels of Sample units				
Groups	G_1	U_1	U_{k_1+1}	.	.	$U_{(s-1)k_1+1}$
	G_2	U_2	U_{k_1+2}	.	.	$U_{(s-1)k_1+2}$
	G_3	U_3	U_{k_1+3}	.	.	$U_{(s-1)k_1+3}$
	G_i	U_i	U_{k_1+i}	.	.	$U_{(s-1)k_1+i}$
	G_{k_1}	U_{k_1}	U_{2k_1}	.	.	$U_{sk_1=N}$

Thus sample units with random starts $r_i (i = 1, 2, ..., m)$ selected from 1 to k_1 correspond to the following labels:

$$r_i + (j-1)k_1, \qquad i = 1, 2, ..., m \quad \text{and} \quad j = 1, 2, ..., s. \tag{2.1}$$

2.1. Estimation of Population Mean and its Variance in MSSM

Consider the mean estimator

$$\bar{y}_{MSSM} = \frac{1}{ms}\sum_{i=1}^{m}\sum_{j=1}^{s} y_{r_ij} = \frac{1}{m}\sum_{i=1}^{m}\Big(\frac{1}{s}\sum_{j=1}^{s} y_{r_ij}\Big).$$

where y_{r_ij} is the value of the jth unit of the ith random group.

Taking expectation on both sides, we get:

$$\begin{aligned} E\left(\bar{y}_{MSSM}\right) &= \frac{1}{m}\sum_{i=1}^{m} E\Big(\frac{1}{s}\sum_{j=1}^{s} y_{r_ij}\Big) \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{k_1}\sum_{i=1}^{k_1}\Big(\frac{1}{s}\sum_{j=1}^{s} y_{ij}\Big) = \frac{1}{sk_1}\sum_{i=1}^{k_1}\sum_{j=1}^{s} y_{ij} = \mu, \end{aligned}$$

where y_{ij} is the value of the jth unit of the ith group and μ is the population mean.

The variance of $\bar{y}_{MSSM}$ is given by

$$V\left(\bar{y}_{MSSM}\right) = E\left(\bar{y}_{MSSM} - \mu\right)^2 = \frac{1}{m^2}\,E\Bigg[\sum_{i=1}^{m}\left(\bar{y}_{r_i\cdot} - \mu\right)\Bigg]^2,$$

where $\bar{y}_{r_i\cdot}$ is the mean of ith random group.

After simplification, we have:

$$V\left(\bar{y}_{MSSM}\right) = \frac{1}{mk_1}\,\frac{(k_1 - m)}{(k_1 - 1)}\,\sum_{i=1}^{k_1}\left(\bar{y}_{i\cdot} - \mu\right)^2, \tag{2.2}$$

where $\bar{y}_{i\cdot}$ is the mean of ith group.

Further, it can be observed that in a situation when MSSM becomes LSS, the variance expression given in Equation (2.2) reduces to variance of LSS, i.e., $V\left(\bar{y}_{MSSM}\right) = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{y}_{i\cdot} - \mu\right)^2 = V\left(\bar{y}_{LSS}\right)$.

Similarly, in the case when MSSM becomes SRS, $V\left(\bar{y}_{MSSM}\right)$ reduces to variance of SRS without replacement, i.e.,

$$V\left(\bar{y}_{MSSM}\right) = \frac{(N-n)}{nN}\,\frac{1}{(N-1)}\,\sum_{i=1}^{N}\left(y_i - \mu\right)^2 = V\left(\bar{y}_{SRSWOR}\right).$$

The alternative expressions for $V\left(\bar{y}_{MSSM}\right)$ have been presented in Theorems 2.1, 2.2 and 2.3:

Theorem 2.1. *The variance of sample mean under MSSM is:*

$$V(\bar{y}_{MSSM}) = \frac{1}{mN}\frac{(k_1-m)}{(k_1-1)}\Big[(N-1)S^2 - k_1(s-1)S_{wg}^2\Big],$$

where $S^2 = \dfrac{1}{N-1}\sum_{i=1}^{k_1}\sum_{j=1}^{s}(y_{ij}-\mu)^2$, *and* $S_{wg}^2 = \dfrac{1}{k_1(s-1)}\sum_{i=1}^{k_1}\sum_{j=1}^{s}(y_{ij}-\bar{y}_i)^2$ *is the variance among the units that lie within the same group.*

Proof: From analysis of variance, we have:

$$\sum_{i=1}^{N}(y_i-\mu)^2 = s\sum_{i=1}^{k_1}(\bar{y}_i-\mu)^2 + \sum_{i=1}^{k_1}\sum_{j=1}^{s}(y_{ij}-\bar{y}_i)^2, \quad \text{or}$$

$$(N-1)S^2 = s\sum_{i=1}^{k_1}(\bar{y}_i-\mu)^2 + k_1(s-1)S_{wg}^2.$$

Thus

$$V(\bar{y}_{MSSM}) = \frac{1}{mN}\frac{(k_1-m)}{(k_1-1)}\Big[(N-1)S^2 - k_1(s-1)S_{wg}^2\Big]. \tag{2.3}$$ □

Theorem 2.2. *The variance of sample mean under MSSM is:*

$$V(\bar{y}_{MSSM}) = \frac{1}{n}\left(\frac{k_1-m}{k_1-1}\right)\left(\frac{N-1}{N}\right)S^2\Big[1+(s-1)\rho_w\Big],$$

where

$$\rho_w = \frac{\sum_{i=1}^{k_1}\sum_{j=1}^{s}\sum_{\substack{j'=1\\ j'\neq j}}^{s}(y_{ij}-\mu)(y_{ij'}-\mu)/s(s-1)k_1}{\sum_{i=1}^{k_1}\sum_{j=1}^{s}(y_{ij}-\mu)^2/sk_1}.$$

Proof: Note that

$$\begin{aligned}
V(\bar{y}_{MSSM}) &= \frac{1}{mk_1}\frac{(k_1-m)}{(k_1-1)}\sum_{i=1}^{k_1}(\bar{y}_i-\mu)^2 \\
&= \frac{1}{s^2mk_1}\frac{(k_1-m)}{(k_1-1)}\sum_{i=1}^{k_1}\Big[\sum_{j=1}^{s}(y_{ij}-\mu)\Big]^2 \\
&= \frac{1}{s^2mk_1}\frac{(k_1-m)}{(k_1-1)}\Big[\sum_{i=1}^{k_1}\sum_{j=1}^{s}(y_{ij}-\mu)^2 + \sum_{i=1}^{k_1}\sum_{j\neq 1}^{s}(y_{ij}-\mu)(y_{iu}-\mu)\Big] \\
&= \frac{1}{s^2mk_1}\frac{(k_1-m)}{(k_1-1)}\Big[(sk_1-1)S^2 + (sk_1-1)(s-1)S^2\rho_w\Big].
\end{aligned}$$

Hence

$$V\left(\bar{y}_{MSSM}\right) = \frac{1}{n}\,\frac{(k_1-m)}{(k_1-1)}\,\frac{(N-1)}{N}\,S^2\Big[1+(s-1)\rho_w\Big], \tag{2.4}$$

where ρ_w is the intraclass correlation between the pairs of units that are in the same group. □

Theorem 2.3. *The variance of $\bar{y}_{MSSM}$ is:*

$$V\left(\bar{y}_{MSSM}\right) = \frac{(k_1-m)}{mN}\,S^2_{wst}\Big[1+(s-1)\rho_{wst}\Big],$$

where

$$S^2_{wst} = \frac{1}{s(k_1-1)}\sum_{j=1}^{s}\sum_{i=1}^{k_1}(y_{ij}-\bar{y}_{.j})^2$$

and

$$\rho_{wst} = \frac{\sum\limits_{i=1}^{k_1}\sum\limits_{j=1}^{s}\sum\limits_{\substack{j'=1\\ j'\neq j}}^{s}(y_{ij}-\bar{y}_j)\,(y_{ij'}-\bar{y}_{j'})}{s(s-1)\,(k_1-1)\,S^2_{wst}}.$$

Proof: Note that

$$\begin{aligned}
V\left(\bar{y}_{MSSM}\right) &= \frac{1}{mk_1}\,\frac{(k_1-m)}{(k_1-1)}\sum_{i=1}^{k_1}(\bar{y}_i-\mu)^2\\
&= \frac{1}{mk_1}\,\frac{(k_1-m)}{(k_1-1)}\sum_{i=1}^{k_1}\Big[\frac{1}{s}\sum_{j=1}^{s}y_{ij}-\frac{1}{s}\sum_{j=1}^{s}\bar{y}_j\Big]^2\\
&= \frac{1}{s^2mk_1}\,\frac{(k_1-m)}{(k_1-1)}\sum_{i=1}^{k_1}\Big[\sum_{j=1}^{s}(y_{ij}-\bar{y}_j)\Big]^2\\
&= \frac{1}{smN}\,\frac{(k_1-m)}{(k_1-1)}\left[\sum_{j=1}^{s}\sum_{i=1}^{k_1}(y_{ij}-\bar{y}_j)^2+\sum_{i=1}^{k_1}\sum_{j=1}^{s}\sum_{\substack{j'=1\\ j'\neq j}}^{s}(y_{ij}-\bar{y}_j)(y_{ij'}-\bar{y}_{j'})\right]\\
&= \frac{1}{smN}\,\frac{(k_1-m)}{(k_1-1)}\,s(k_1-1)\,S^2_{wst}\Big[1+(s-1)\rho_{wst}\Big].
\end{aligned}$$

Hence

$$V\left(\bar{y}_{MSSM}\right) = \left(\frac{k_1-m}{mN}\right)S^2_{wst}\Big[1+(s-1)\rho_{wst}\Big]. \tag{2.5}$$

□

3. MEAN, VARIANCE AND EFFICIENCY COMPARISON OF MSSM FOR POPULATIONS EXHIBITING LINEAR TREND

Generally the efficiency of every new systematic sampling design is evaluated for populations having linear trend. Therefore, consider the following linear model for the hypothetical population

$$Y_t = \alpha + \beta t, \qquad t = 1, 2, 3, ..., N, \tag{3.1}$$

where α and β respectively are the intercept and slope terms in the model.

3.1. Sample Mean under MSSM

$$\bar{y}_{MSSM} = \alpha + \frac{\beta}{ms}\sum_{i=1}^{m}\sum_{j=1}^{s}\Big\{r_i + (j-1)k_1\Big\}, \quad \text{or}$$

$$\bar{y}_{MSSM} = \alpha + \frac{\beta}{m}\Big\{\sum_{i=1}^{m} r_i + \frac{m}{2}(s-1)k_1\Big\}. \tag{3.2}$$

$$E\left(\bar{y}_{MSSM}\right) = \alpha + \beta\frac{(N+1)}{2} = \mu. \tag{3.3}$$

$$V\left(\bar{y}_{MSSM}\right) = E\left\{\bar{y}_{MSSM} - E(\bar{y}_{MSSM})\right\}^2 = \beta^2 E\Big[\frac{1}{m}\sum_{i=1}^{m} r_i - \frac{(k_1+1)}{2}\Big]^2.$$

Hence

$$V\left(\bar{y}_{MSSM}\right) = \beta^2\frac{(k_1+1)(k_1-m)}{12m}. \tag{3.4}$$

Note that $m = 1$ and $k_1 = k$ in situations when MSSM is LSS; therefore

$$V\left(\bar{y}_{MSSM}\right) = \beta^2\frac{(k^2-1)}{12} = V\left(\bar{y}_{LSS}\right). \tag{3.5}$$

Similarly, $m = n$ and $k_1 = N$ in situations when MSSM is SRS, so

$$V\left(\bar{y}_{MSSM}\right) = \beta^2\frac{(N+1)(N-n)}{12n} = V\left(\bar{y}_{SRS}\right). \tag{3.6}$$

The efficiency of MSSM with respect to SRS can be calculated as below:

$$\textit{Efficiency} = \frac{V\left(\bar{y}_{SRS}\right)}{V\left(\bar{y}_{MSSM}\right)} = \frac{m(N+1)(N-n)}{(k_1+1)(k_1-m)n} = \frac{(sk_1+1)}{(k_1+1)} \geq 1, \tag{3.7}$$

as $s \geq 1$. One can see that MSSM is always more efficient than SRS if $s > 1$ and is equally efficient if $s = 1$.

4. ESTIMATION OF VARIANCE

Sampath and Ammani (2012) have considered LSS, Balanced Systematic Sampling (BSS) proposed by Sethi (1965), and Modified Systematic Sampling (MS) proposed by Singh *et al.* (1968) using multiple random starts. They have derived excellent expressions of unbiased variance estimators and their variances for these schemes. However, these schemes are not applicable if $N \neq nk$. Fortunately, MSSM nicely handles this by producing unbiased variance estimator and its variance for the case, where $N \neq nk$. Adopting the procedure mentioned in Sampath and Ammani (2012), we can get an unbiased variance estimator and its variance in MSSM for the case where $N \neq nk$.

In MSSM, the probability that the i^{th} unit will be included in the sample is just the probability of including the group containing the specific unit in the sample. Hence, the first-order inclusion probability that corresponds to the population unit with label i, is given by

$$\pi_i = \frac{m}{k_1} = \frac{ms}{sk_1} = \frac{n}{N}, \qquad i = 1, 2, 3, ..., N.$$

In the second-order inclusion probabilities, the pairs of units may belong to the same or the different groups. The pairs of units belong to the same group only if the respective group is included in the sample. Thus, the second-order inclusion probabilities for pairs of units belonging to the same group are equivalent to the first-order inclusion probabilities, i.e.,

$$\pi_{ij} = \frac{m}{k_1} = \frac{ms}{sk_1} = \frac{n}{N}, \qquad i,\, j \in s_{r_u} \text{ for some } r_u \quad (r_u = 1, 2, ..., k_1).$$

On the other hand, pairs of units belonging to two different groups occurs only when the corresponding pair of groups is included in the sample. Hence, the second-order inclusion probability is given by

$$\pi_{ij} = \frac{m(m-1)}{k_1(k_1-1)}, \qquad \text{if } i \in s_{r_u} \text{ and } j \in s_{r_v} \text{ for some } u \neq v.$$

Thus

$$\pi_{ij} = \frac{m(m-1)}{k_1(k_1-1)} = \frac{ms(ms-s)}{sk_1(sk_1-s)} = \frac{n(n-s)}{N(N-s)}.$$

Since the second-order inclusion probabilities are positive for all pairs of units in the population, an unbiased estimator of population variance can be established. To accomplish this, the population variance

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu)^2$$

can be written as

$$S^2 = \frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} (Y_i - Y_j)^2.$$

By using second-order inclusion probabilities, an unbiased estimator of the population variance can be obtained as

$$\hat{S}^2_{MSSM} = \frac{1}{2N(N-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{(y_i - y_j)^2}{\pi_{ij}}.$$

As $n = ms$, it means that there are m random sets each containing s units. Therefore, taking r_u $(u = 1, 2, ...m)$ as the random start for the u^{th} set, the expression for $\hat{S}^2_{MSSM}$ can be rewritten as:

$$\begin{aligned}
\hat{S}^2_{MSSM} &= \frac{1}{2N(N-1)} \left[\sum_{u=1}^{m} \left\{ \sum_{i=1}^{s} \sum_{\substack{j=1 \\ j\neq i}}^{s} \frac{(y_{r_u i} - y_{r_u j})^2}{\pi_{ij}} \right\} \right. \\
&\quad \left. + \sum_{\substack{u=1 \\ v=1 \\ u\neq v}}^{m} \left\{ \sum_{i=1}^{s} \sum_{j=1}^{s} \frac{(y_{r_u i} - y_{r_v j})^2}{\pi_{ij}} \right\} \right] \\
&= \frac{1}{2N(N-1)} \left[\frac{N}{n} \sum_{u=1}^{m} \left\{ \sum_{i=1}^{s} \sum_{\substack{j=1 \\ j\neq i}}^{s} (y_{r_u i} - y_{r_u j})^2 \right\} \right. \\
&\quad \left. + \frac{N(N-s)}{n(n-s)} \sum_{\substack{u=1 \\ v=1 \\ u\neq v}}^{m} \left\{ \sum_{i=1}^{s} \sum_{j=1}^{s} (y_{r_u i} - y_{r_v j})^2 \right\} \right] \\
&= \frac{1}{2N(N-1)} \left[\frac{N}{n} \sum_{u=1}^{m} \left\{ 2s \sum_{i=1}^{s} (y_{r_u i} - \bar{y}_{r_u})^2 \right\} \right. \\
&\quad \left. + \frac{N(N-s)}{n(n-s)} \sum_{\substack{u=1 \\ v=1 \\ u\neq v}}^{m} \left\{ \sum_{i=1}^{s} \sum_{j=1}^{s} \left((y_{r_u i} - \bar{y}_u)^2 + (y_{r_v j} - \bar{y}_{r_v})^2 + (\bar{y}_{r_u} - \bar{y}_{r_v})^2 \right) \right\} \right] \\
&= \frac{1}{2N(N-1)} \left[\frac{N}{n} \sum_{u=1}^{m} \left\{ 2s^2 \hat{\sigma}^2_{r_u} \right\} \right. \\
&\quad \left. + \frac{N(N-s)}{n(n-s)} \sum_{\substack{u=1 \\ v=1 \\ u\neq v}}^{m} \left\{ \left(s^2 \hat{\sigma}^2_{r_u} + s^2 \hat{\sigma}^2_{r_v} + s^2 (\bar{y}_{r_u} - \bar{y}_{r_v})^2 \right) \right\} \right],
\end{aligned}$$

where $\bar{y}_{r_u}$ and $\hat{\sigma}^2_{r_u} = \frac{1}{s} \sum_{i=1}^{s} (y_{r_u i} - \bar{y}_{r_u})^2$ are the mean and variance of the u^{th} group

$(u = 1, 2, ..., m)$. Further,

$$\begin{aligned}
\hat{S}^2_{MSSM} &= \frac{1}{2N(N-1)}\left[\frac{N}{n}\Big\{2s^2\sum_{u=1}^{m}\hat{\sigma}^2_{r_u}\Big\}\right. \\
&\quad \left. + \frac{N(N-s)}{n(n-s)}\Big\{\Big(2(m-1)s^2\sum_{u=1}^{m}\hat{\sigma}^2_{r_u} + s^2\sum_{u=1}^{m}\sum_{\substack{v=1\\u\neq v}}^{m}(\bar{y}_{r_u}-\bar{y}_{r_v})^2\Big)\Big\}\right] \\
&= \frac{1}{2N(N-1)}\left[\frac{N}{n}\Big\{2s^2\sum_{u=1}^{m}\hat{\sigma}^2_{r_u}\Big\}\right. \\
&\quad \left. + \frac{N(N-s)}{n(n-s)}\Big\{\Big(2(m-1)s^2\sum_{u=1}^{m}\hat{\sigma}^2_{r_u} + s^2 2\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}(\bar{y}_{r_u}-\bar{y}_{r_v})^2\Big)\Big\}\right] \\
&= \frac{s^2}{ms(N-1)}\left[\sum_{u=1}^{m}\hat{\sigma}^2_{r_u}\left\{1+\frac{(N-s)}{(ms-s)}(m-1)\right\}\right. \\
&\quad \left. + \frac{(N-s)}{(ms-s)}\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}(\bar{y}_{r_u}-\bar{y}_{r_v})^2\right].
\end{aligned}$$

Hence

$$\hat{S}^2_{MSSM} = \frac{1}{(N-1)}\left[\sum_{u=1}^{m}\hat{\sigma}^2_{r_u}\frac{N}{m} + \frac{(N-s)}{m(m-1)}\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}(\bar{y}_{r_u}-\bar{y}_{r_v})^2\right]. \tag{4.1}$$

For simplicity, Equation (4.1) can be written as

$$\hat{S}^2_{MSSM} = \frac{1}{(N-1)}\left[\frac{N}{m}\sum_{u=1}^{m}\hat{\sigma}^2_{r_u} + \frac{(N-s)}{(m-1)}\sum_{u=1}^{m}(\bar{y}_{r_u}-\bar{y}_{MSSM})^2\right].$$

The resulting estimator obtained in Equation (4.1) is an unbiased estimator of population variance S^2. It is mentioned in Section 2, if lcm of N and n is the product of N and n, i.e., $L = N \times n$, then MSSM becomes SRS.

Consequently, $\hat{\sigma}^2_{r_u} = 0 (u = 1, 2, ..., m)$ and

$$\hat{S}^2_{MSSM} = \hat{S}^2_{SRS} = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

which is a well-known unbiased estimator of S^2 in SRS without replacement.

4.1. Variance of $\hat{S}^2_{MSSM}$

The variance of $\hat{S}^2_{MSSM}$ is given by

(4.2)
$$V\left(\hat{S}^2_{MSSM}\right) = \frac{1}{m\left(N-1\right)^2}\left[\frac{N^2\left(k_1-m\right)}{\left(k_1-1\right)}\,\sigma_0^2 + \frac{\left(N-s\right)^2 k_1}{\left(m-1\right)}\left[\left\{\frac{\left(m-1\right)}{\left(k_1-1\right)} - \frac{\left(m-2\right)\left(m-3\right)}{\left(k_1-2\right)\left(k_1-3\right)}\right\}\mu_4 + \left\{\frac{\left(k_1-3\right)-\left(m-2\right)\left(k_1+3\right)}{\left(k_1-1\right)^2} + \frac{\left(m-2\right)\left(m-3\right)\left(k_1^2-3\right)}{\left(k_1-1\right)^2\left(k_1-2\right)\left(k_1-3\right)}\right\}\mu_2^2\right] + 2\,\frac{N\left(N-s\right)\left(k_1-m\right)}{\left(k_1-1\right)\left(k_1-2\right)}\left\{\sum_{r=1}^{k_1}\hat{\sigma}_r^2\left(\bar{y}_r-\bar{Y}\right)^2 - k_1\bar{\sigma}^2\mu_2\right\}\right]$$

(see details in Appendix A).

Note that, if $L = N$, then MSSM becomes LSS and the above formula is not valid in this case. Fortunately, in LSS, due to Gautschi (1957), the population is divided into $m'k$ groups of n/m' elements, and m' of these groups will randomly be selected to get a sample of size n. Thus, one can easily modify the above formula by just putting $m = m'$, $k_1 = m'k$ and $s = n/m'$ in Equation (A.9) and get $V\left(\hat{S}^2_{LSS}\right)$ as below:

(4.3)
$$V\left(\hat{S}^2_{LSS}\right) = \frac{1}{m'\left(N-1\right)^2}\left[\frac{N^2 m'\left(k-1\right)}{\left(m'k-1\right)}\,\sigma_0^2 + \frac{\left(m'N-n\right)^2 k}{m'\left(m'-1\right)}\left[\left\{\frac{\left(m'-1\right)}{\left(m'k-1\right)} - \frac{\left(m'-2\right)\left(m'-3\right)}{\left(m'k-2\right)\left(m'k-3\right)}\right\}\mu_4 + \left\{\frac{\left(m'k-3\right)-\left(m'-2\right)\left(m'k+3\right)}{\left(m'k-1\right)^2} + \frac{\left(m'-2\right)\left(m'-3\right)\left(m'^2k^2-3\right)}{\left(m'k-1\right)^2\left(m'k-2\right)\left(m'k-3\right)}\right\}\mu_2^2\right] + 2\,\frac{N\left(m'N-n\right)\left(k-1\right)}{\left(m'k-1\right)\left(m'k-2\right)}\left\{\sum_{r=1}^{m'k}\hat{\sigma}_r^2\left(\bar{y}_r-\mu\right)^2 - m'k\bar{\sigma}^2\mu_2\right\}\right].$$

This is the general formula for the variance of unbiased variance estimator with m' random starts for LSS. Further, one can also easily deduce the following

formula of $V\left(\hat{S}^2_{SRS}\right)$ by putting $k_1 = N$, $m = n$ and $s = 1$ in Equation (A.9):

$$
\begin{aligned}
V\left(\hat{S}^2_{SRS}\right) = \frac{N}{n\,(n-1)} \Bigg[& \left\{ \frac{(n-1)}{(N-1)} - \frac{(n-2)\,(n-3)}{(N-2)\,(N-3)} \right\} \mu_4 \\
& + \left\{ \frac{(N-3) - (n-2)\,(N+3)}{(N-1)^2} \right. \\
& \left. + \frac{\left(N^2-3\right)(n-2)\,(n-3)}{(N-1)^2\,(N-2)\,(N-3)} \right\} \mu_2^2 \Bigg].
\end{aligned} \tag{4.4}
$$

5. EFFICIENCY COMPARISON OF VARIANCE ESTIMATORS

In this section, we compare $\hat{S}^2_{MSSM}$ with $\hat{S}^2_{SRS}$ by using natural and simulated populations. Furthermore, this study is carried out for those choices of sample sizes in which the condition "$N < L < (N \times n)$" is satisfied. It has already been mentioned that MSSM becomes LSS when $L = N$. On the other hand, MSSM becomes SRS when $L = (N \times n)$.

5.1. Natural Populations

In Population 1 (see Murthy, 1967, p. 131–132), the data on volume of timber of 176 forest strips have been considered. In this data, the volume of timber has been arranged with respect to its length. In Population 2 (see Murthy, 1967, p. 228), the data of output along with the fixed capital of 80 factories have been considered. Here, output is arranged with respect to fixed capital. It is observed that the data considered in Population 1 and Population 2 approximately follow a linear trend. In this empirical study, the variances of $\hat{S}^2_{MSSM}$ and $\hat{S}^2_{SRS}$ are computed for various sample sizes and efficiency is computed using the expression:

$$
\textit{Efficiency} = \frac{V\left(\hat{S}^2_{SRS}\right)}{V\left(\hat{S}^2_{MSSM}\right)}.
$$

The population size N, sample size n, number of random starts m, number of elements in each group s, the number of groups k_1 containing the N units of the population and efficiency of MSSM over SRS are respectively presented in Columns 1 to 6 for Population 1 and Columns 7 to 12 for Population 2 in Table 2. From the efficiency comparison presented in Table 2, it has been observed that MSSM is more efficient than SRS. Moreover, one can also see that as the number of elements s in each group are increased, the efficiency of MSSM also increases. Such increase in efficiency is due to the fact that in MSSM, the units

within the groups are arranged in a systematic pattern. So, more number of units with systematic pattern will cause increase in efficiency.

Table 2: Efficiency comparison of $\hat{S}^2_{MSSM}$ and $\hat{S}^2_{SRS}$ in both natural populations.

Population 1						Population 2					
N	n	m	s	k_1	Efficiency	N	n	m	s	k_1	Efficiency
176	10	5	2	88	1.41	80	6	3	2	40	2.31
	12	3	4	44	3.69		12	3	4	20	3.56
	14	7	2	88	2.04		14	7	2	40	2.29
	18	9	2	88	2.03		15	3	5	16	5.91
	20	5	4	44	3.64		18	9	2	40	2.28
	24	3	8	22	5.79		22	11	2	40	2.27
	26	13	2	88	2.01		24	3	8	10	14.11
	28	7	4	44	3.61		25	5	5	16	5.91
	30	15	2	88	2.00		26	13	2	40	2.27
	32	2	16	11	6.22		28	7	4	20	3.50
	34	17	2	88	2.00		30	3	10	8	10.85
	36	9	4	44	3.59		32	2	16	5	15.14
	38	19	2	88	2.00		34	17	2	40	2.26
	40	5	8	22	5.70		35	7	5	16	5.89
	42	21	2	88	1.99		36	9	4	20	3.49
	46	23	2	88	1.99		38	19	2	40	2.26
	50	25	2	88	1.99						

5.2. Simulated Populations

The simulation study, two populations of sizes 160 and 280 are generated for the following distribution with variety of parameters by using R-packages:

(**i**) *Uniform distribution:* Here only three sets of the parametric values are considered, i.e., (10, 20), (10, 30) and (10, 50).

(**ii**) *Normal distribution:* In this case, six sets of parametric values are considered with means 20, 40 and 60 and standard deviations 5 and 8.

(**iii**) *Gamma distribution:* Eight sets of parametric values are considered in this case. Here, 1, 3, 5 and 10 are considered as the values of scale parameter with 2 and 4 as the values of shape parameter.

In each distribution, using each combination of the parametric values for each choice of the sample size, each population with and without order is replicated 1000 times. $V\left(\hat{S}^2_{MSSM}\right)$ and $V\left(\hat{S}^2_{SRS}\right)$ are computed for each population (with and without order) for the various choices of sample sizes. The average of 1000 values of the variances of $\hat{S}^2_{MSSM}$ and $\hat{S}^2_{SRS}$ is then computed for each population.

The efficiencies, *Eff*1 and *Eff*2 of MSSM compared to SRS are computed using the following expressions:

$$Eff1 = \frac{Average\{V(\hat{S}^2_{SRSWOR})\}}{Average\{V(\hat{S}^2_{MSSM})\}} \qquad \text{without ordered population}$$

and

$$Eff2 = \frac{Average\{V(\hat{S}^2_{SRSWOR})\}}{Average\{V(\hat{S}^2_{MSSM})\}} \qquad \text{with ordered population.}$$

The efficiencies, *Eff*1 and *Eff*2 for Uniform distribution, Normal distribution and Gamma distribution are presented in Tables 3, 4 and 5 respectively.

It is observed that *Eff*1 is approximately equal to 1 for almost all choices of parametric values and sample sizes. This mean that MSSM and SRS are equally efficient in case of random populations. Thus, for such populations, MSSM can be preferred over SRS due to the qualities that there are no more issues of unbiased estimation of population variance.

Furthermore, it is also observed from Tables 3, 4 and 5 that *Eff*2 is greater than 1 in all cases. It indicates that MSSM is more efficient than SRS in ordered populations. The discussion of *Eff*2 in Tables 3, 4 and 5 is as follows:

In Table 3, the efficiency (*Eff*2) is not effected much by the different combinations of parametric values of the uniform distribution and changes are caused by the number of groups k_1. It is also observed that MSSM is much more efficient for the ordered populations of uniform distribution as compared to the normal and gamma distributions.

In Table 4, the efficiency *Eff*2 is also not much changed like uniform distribution for different combinations of parametric values of the normal distribution. However, *Eff*2 is mainly changed due to the formation of number of groups k_1 of the population units in MSSM. Efficiency will increase with the decrease in the number of groups k_1, and it will decrease with the increase in the number of groups k_1.

In Table 5, the efficiency *Eff*2 is effected by the number of groups k_1 along with the shape parameter of the Gamma distribution. However, change in scale parameter has no significant effect on efficiency of MSSM. Here also the efficiency increases with decrease in the number of groups k_1.

From the above discussion, it is obvious that MSSM performs better than SRS for the populations that follow uniform and parabolic trends. However, such populations must be ordered with certain characteristics. To know further about the performance of MSSM, it would be interesting to study the variances of $\hat{S}^2_{MSSM}$ and $\hat{S}^2_{SRS}$ in the presence of linear trend. This study has been carried out in the following section.

Table 3: Efficiency of MSSM over SRS using uniform distribution.

Uniform Distribution										
N	n	m	s	k_1	$a=10,\ b=20$		$a=10,\ b=30$		$a=10,\ b=50$	
					Eff 1	*Eff* 2	*Eff* 1	*Eff* 2	*Eff* 1	*Eff* 2
160	12	3	4	40	0.94	24.17	0.94	24.67	0.94	24.33
	14	7	2	80	1.00	5.99	0.99	6.04	0.98	6.05
	15	3	5	32	0.95	39.72	0.96	38.39	0.95	39.85
	18	9	2	80	0.99	6.19	0.99	6.16	1.00	6.15
	22	11	2	80	1.00	6.17	1.00	6.26	1.00	6.25
	24	3	8	20	0.94	91.20	0.95	90.56	0.98	89.22
	25	5	5	32	0.99	44.66	0.99	44.24	0.98	45.01
	26	13	2	80	1.00	6.30	1.00	6.37	1.00	6.26
	28	7	4	40	0.99	29.47	1.00	30.36	0.99	28.90
	30	3	10	16	0.98	125.89	0.98	125.98	0.96	120.79
	34	17	2	80	1.00	6.33	1.00	6.50	1.00	6.44
	35	7	5	32	1.00	43.94	0.99	46.98	0.99	45.59
	36	9	4	40	0.99	30.19	1.00	30.60	0.99	30.23
	38	19	2	80	1.00	6.44	1.00	6.45	0.99	6.37
280	12	3	4	70	0.94	28.07	0.94	28.44	0.93	28.48
	15	3	5	56	0.94	47.65	0.94	47.37	0.94	47.68
	16	2	8	35	0.89	102.47	0.89	104.25	0.90	103.83
	18	9	2	140	0.99	6.41	0.99	6.50	0.99	6.51
	22	11	2	140	0.99	6.61	0.99	6.48	1.00	6.60
	24	3	8	35	0.96	123.18	0.96	121.88	0.96	124.69
	25	5	5	56	0.98	55.48	0.98	56.05	0.99	54.97
	26	13	2	140	0.99	6.74	0.99	6.77	1.00	6.62
	30	3	10	28	0.96	189.18	0.97	182.58	0.97	184.39
	32	4	8	35	0.97	135.94	0.98	131.74	0.99	134.95
	34	17	2	140	1.00	6.75	1.00	6.88	1.00	6.84
	36	9	4	70	0.99	38.04	1.00	36.47	0.99	36.59
	38	19	2	140	1.00	6.82	1.00	6.85	1.00	6.83
	42	3	14	20	0.99	292.91	0.98	320.06	0.96	310.45
	44	11	4	70	1.00	38.00	0.99	37.56	1.00	37.28
	45	9	5	56	1.00	61.45	1.00	60.35	1.00	59.46
	46	23	2	140	1.00	7.02	1.00	6.86	1.00	6.95
	48	6	8	35	0.99	144.63	0.99	148.64	1.01	141.89
	49	7	7	40	1.00	111.72	1.00	118.32	1.00	114.09
	50	5	10	28	0.99	195.80	0.99	199.00	0.99	207.92

Table 4: Efficiency of MSSM over SRS using normal distribution.

Normal distribution																
N	n	m	s	k_1	$\sigma = 5$						$\sigma = 10$					
					$\mu = 20$		$\mu = 40$		$\mu = 60$		$\mu = 20$		$\mu = 40$		$\mu = 60$	
					Eff 1	*Eff* 2	*Eff* 1	*Eff* 2	*Eff* 1	*Eff* 2	*Eff* 1	*Eff* 2	*Eff1*	*Eff* 2	*Eff* 1	*Eff* 2
160	12	3	4	40	0.97	3.33	0.97	3.34	0.98	3.35	0.96	3.34	0.97	3.31	0.97	3.28
	14	7	2	80	0.99	1.79	1.00	1.80	1.00	1.79	1.01	1.79	0.99	1.80	1.00	1.80
	15	3	5	32	0.98	4.01	0.97	4.11	0.98	4.11	0.97	4.00	0.97	4.02	0.99	4.06
	18	9	2	80	1.00	1.78	0.99	1.78	0.99	1.80	0.99	1.79	1.00	1.78	1.00	1.79
	22	11	2	80	1.00	1.79	0.99	1.80	0.99	1.78	1.00	1.79	1.00	1.78	0.99	1.77
	24	3	8	20	0.98	6.04	0.98	6.33	0.98	6.19	0.98	6.23	1.00	6.19	0.97	6.16
	25	5	5	32	0.99	3.98	0.99	4.04	0.99	4.05	0.98	4.06	0.99	4.07	1.00	4.03
	26	13	2	80	1.00	1.79	1.00	1.79	1.00	1.77	1.00	1.78	1.00	1.79	0.99	1.79
	28	7	4	40	0.99	3.30	0.99	3.28	1.00	3.27	1.00	3.32	0.99	3.30	0.99	3.32
	30	3	10	16	0.98	7.49	0.98	7.67	0.99	7.72	0.99	7.54	1.00	7.41	0.99	7.77
	34	17	2	80	1.00	1.78	1.00	1.78	1.00	1.79	0.99	1.79	1.00	1.79	1.00	1.78
	35	7	5	32	1.00	4.00	1.00	4.03	0.99	4.03	0.99	4.04	0.99	3.99	1.01	4.01
	36	9	4	40	1.00	3.34	1.00	3.25	1.00	3.28	1.01	3.32	0.99	3.30	1.00	3.29
	38	19	2	80	0.99	1.81	1.00	1.79	1.00	1.77	1.00	1.78	1.00	1.78	1.00	1.79
280	12	3	4	70	0.96	3.34	0.97	3.31	0.96	3.33	0.97	3.33	0.97	3.35	0.97	3.33
	15	3	5	56	0.96	4.12	0.98	4.03	0.97	4.05	0.97	4.14	0.99	4.09	0.97	4.01
	16	2	8	35	0.95	6.30	0.95	6.16	0.94	6.26	0.95	6.31	0.95	6.29	0.94	6.35
	18	9	2	140	1.00	1.79	0.99	1.79	1.00	1.80	0.99	1.79	1.00	1.79	1.00	1.79
	22	11	2	140	1.00	1.79	1.00	1.78	1.00	1.80	0.99	1.79	1.00	1.80	1.00	1.80
	24	3	8	35	0.98	6.17	0.99	6.35	0.99	6.06	0.98	6.25	0.98	6.14	0.98	6.26
	25	5	5	56	0.99	4.01	1.00	4.11	1.00	4.05	1.00	4.07	0.99	4.04	0.99	4.02
	26	13	2	140	1.00	1.80	0.99	1.79	1.00	1.80	1.00	1.78	1.00	1.81	1.00	1.79
	30	3	10	28	0.98	7.53	0.98	7.73	0.99	7.72	0.99	7.98	1.00	7.78	0.99	7.71
	32	4	8	35	0.99	6.17	1.00	6.38	1.00	6.16	1.00	6.35	0.99	6.16	0.99	6.37
	34	17	2	140	1.00	1.78	1.00	1.79	0.99	1.78	1.00	1.78	1.00	1.79	1.00	1.79
	36	9	4	70	1.00	3.33	0.99	3.33	1.00	3.33	0.99	3.28	1.00	3.34	0.99	3.38
	38	19	2	140	1.00	1.78	1.00	1.78	1.00	1.79	1.00	1.79	1.00	1.79	1.00	1.80
	42	3	14	20	0.98	10.32	0.98	10.12	0.99	10.35	1.00	10.55	1.00	10.36	0.99	10.53
	44	11	4	70	1.00	3.30	1.00	3.31	1.00	3.36	1.00	3.28	0.99	3.30	1.00	3.30
	45	9	5	56	0.99	4.07	0.99	4.08	1.01	4.01	1.00	4.07	0.99	4.03	1.00	4.10
	46	23	2	140	1.00	1.78	1.00	1.79	0.99	1.79	1.00	1.79	1.00	1.79	0.99	1.78
	48	6	8	35	0.99	6.22	0.98	6.24	1.00	6.16	0.99	6.19	1.01	6.39	1.00	6.21
	49	7	7	40	1.00	5.66	1.00	5.52	1.00	5.49	0.99	5.49	0.99	5.48	1.00	5.50
	50	5	10	28	1.00	7.89	0.99	7.54	1.00	7.82	0.99	7.59	0.99	7.54	1.01	7.72

Table 5: Efficiency of MSSM over SRS using gamma distribution.

Gamma distribution																
N	n	m	s	k_1	$shape = 2$								$shape = 4$							
					$scale = 1$		$scale = 3$		$scale = 5$		$scale = 10$		$scale = 1$		$scale = 3$		$scale = 5$		$scale = 10$	
					$Eff\,1$	$Eff\,2$	$Eff\,1$	$Eff\,2$	$Eff\,1$	$Eff\,2$	$Eff\,1$	$Eff\,2$	$Eff\,1$	$Eff\,2$	$Eff\,1$	$Eff\,2$	$Eff\,1$	$Eff\,2$	$Eff\,1$	$Eff\,2$
160	12	3	4	40	1.00	1.50	0.98	1.48	0.99	1.45	0.98	1.42	0.98	1.77	0.99	1.75	0.97	1.77	0.99	1.74
	14	7	2	80	0.99	1.21	0.99	1.20	1.00	1.21	1.00	1.21	1.00	1.34	0.99	1.34	1.00	1.33	1.00	1.34
	15	3	5	32	1.01	1.54	0.98	1.57	0.99	1.54	1.00	1.59	0.99	1.89	1.00	1.89	0.97	1.90	0.98	1.89
	18	9	2	80	0.99	1.21	0.99	1.20	1.00	1.21	1.00	1.20	1.00	1.33	1.00	1.34	1.00	1.33	1.00	1.34
	22	11	2	80	1.01	1.20	1.01	1.21	1.00	1.21	1.00	1.20	0.99	1.32	0.99	1.33	1.00	1.32	1.00	1.33
	24	3	8	20	0.99	1.81	1.00	1.79	1.00	1.77	1.00	1.86	0.99	2.32	1.00	2.27	1.00	2.25	1.00	2.26
	25	5	5	32	1.01	1.55	0.99	1.57	1.00	1.53	0.99	1.56	0.99	1.92	1.00	1.87	0.99	1.89	1.00	1.87
	26	13	2	80	1.00	1.20	1.00	1.19	1.00	1.20	1.00	1.20	0.99	1.33	1.00	1.32	1.00	1.33	1.00	1.32
	28	7	4	40	1.00	1.45	1.00	1.44	1.00	1.46	0.99	1.44	0.99	1.76	1.00	1.72	1.00	1.73	0.99	1.72
	30	3	10	16	0.99	1.91	1.00	1.98	1.00	1.90	0.98	1.98	0.98	2.53	0.98	2.44	0.99	2.51	1.01	2.44
	34	17	2	80	1.00	1.20	1.00	1.21	1.00	1.20	1.00	1.19	1.00	1.33	1.00	1.33	1.00	1.33	1.00	1.31
	35	7	5	32	0.99	1.55	0.99	1.56	1.00	1.53	1.01	1.56	1.00	1.86	1.00	1.89	0.98	1.89	1.00	1.89
	36	9	4	40	0.99	1.43	1.00	1.45	1.00	1.47	1.01	1.45	1.00	1.75	0.99	1.76	1.00	1.76	0.99	1.76
	38	19	2	80	0.99	1.20	1.00	1.21	0.99	1.20	1.00	1.20	1.01	1.32	1.00	1.32	1.00	1.32	1.00	1.33
280	12	3	4	70	0.98	1.48	0.99	1.46	0.99	1.46	0.99	1.46	0.98	1.76	0.98	1.78	0.99	1.76	0.99	1.76
	15	3	5	56	0.99	1.57	0.98	1.58	0.99	1.56	1.00	1.57	0.98	1.92	0.98	1.95	0.99	1.89	0.99	1.94
	16	2	8	35	0.99	1.82	0.98	1.80	0.98	1.77	0.97	1.85	0.98	2.29	0.98	2.26	0.98	2.32	0.96	2.29
	18	9	2	140	1.00	1.21	1.00	1.20	1.00	1.21	1.00	1.20	1.00	1.33	1.00	1.33	1.00	1.34	1.00	1.34
	22	11	2	140	1.00	1.21	0.99	1.21	1.00	1.20	1.00	1.21	1.00	1.33	1.00	1.32	1.00	1.32	1.00	1.34
	24	3	8	35	0.98	1.84	0.99	1.79	0.99	1.82	1.00	1.82	0.99	2.31	0.98	2.27	0.99	2.29	0.99	2.24
	25	5	5	56	0.99	1.55	1.01	1.56	1.00	1.53	1.00	1.55	0.99	1.92	1.00	1.88	1.00	1.89	1.00	1.87
	26	13	2	140	1.00	1.20	0.99	1.20	1.00	1.20	1.00	1.21	1.00	1.33	1.00	1.32	1.00	1.32	1.00	1.32
	30	3	10	28	1.00	1.93	0.99	1.91	0.99	1.89	1.00	1.96	0.99	2.48	0.98	2.53	0.98	2.52	0.99	2.47
	32	4	8	35	1.00	1.81	0.99	1.79	1.00	1.80	1.00	1.80	1.00	2.24	1.00	2.27	1.01	2.28	0.98	2.26
	34	17	2	140	1.00	1.20	1.00	1.20	1.00	1.21	1.00	1.19	1.00	1.33	1.00	1.33	1.00	1.33	1.00	1.32
	36	9	4	70	0.99	1.44	1.00	1.44	1.00	1.44	0.99	1.44	0.99	1.71	1.00	1.75	1.00	1.75	1.00	1.70
	38	19	2	140	1.00	1.20	1.00	1.20	1.00	1.19	1.00	1.20	1.00	1.34	1.00	1.32	1.01	1.32	1.00	1.33
	42	3	14	20	0.97	2.13	0.98	2.19	0.99	2.19	0.98	2.13	1.00	2.81	0.98	2.93	1.00	2.83	0.99	2.80
	44	11	4	70	1.00	1.46	1.00	1.46	1.00	1.47	0.99	1.46	0.99	1.71	1.00	1.74	1.00	1.72	1.01	1.74
	45	9	5	56	1.01	1.56	1.00	1.55	1.00	1.54	1.00	1.53	0.99	1.90	1.01	1.91	0.99	1.88	0.98	1.90
	46	23	2	140	1.00	1.20	1.00	1.19	1.00	1.20	1.00	1.20	1.00	1.33	1.00	1.33	1.00	1.32	1.00	1.32
	48	6	8	35	1.01	1.79	1.01	1.82	0.98	1.76	1.01	1.78	1.00	2.27	1.00	2.30	1.00	2.29	1.00	2.26
	49	7	7	40	1.00	1.72	0.99	1.70	0.99	1.70	0.99	1.74	1.00	2.13	1.00	2.16	0.99	2.11	1.00	2.12
	50	5	10	28	1.00	1.92	0.99	1.96	1.00	1.96	0.98	1.94	1.01	2.47	0.99	2.48	1.00	2.46	1.00	2.42

6. VARIANCE OF $\hat{S}^2_{MSSM}$ IN THE PRESENCE OF LINEAR TREND

The variance of $\hat{S}^2_{MSSM}$ under the linear Model (3.1) is given by

(6.1)
$$V\left(\hat{S}^2_{MSSM}\right) = \frac{\beta^4\left(k_1^2-1\right)}{m\left(N-1\right)^2}\frac{\left(N-s\right)^2 k_1}{\left(m-1\right)} \times\left[\frac{\left(3k_1^2-7\right)}{240}\left\{\frac{\left(m-1\right)}{\left(k_1-1\right)}-\frac{\left(m-2\right)\left(m-3\right)}{\left(k_1-2\right)\left(k_1-3\right)}\right\}+\frac{1}{144}\left(k_1^2-1\right) \times\left\{\frac{\left(k_1-3\right)-\left(m-2\right)\left(k_1+3\right)}{\left(k_1-1\right)^2}\frac{\left(m-2\right)\left(m-3\right)\left(k_1^2-3\right)}{\left(k_1-1\right)^2\left(k_1-2\right)\left(k_1-3\right)}\right\}\right]$$

(see details in Appendix B).

Substituting $m = n$, $s = 1$ and $k_1 = N$ in (B.7), the variance of $\hat{S}^2_{SRS}$ can be obtained in the presence of linear trend, i.e.,

(6.2)
$$V\left(\hat{S}^2_{SRS}\right) = \frac{\beta^4\left(N^2-1\right)N}{n(n-1)}\left[\frac{\left(3N^2-7\right)}{240}\left\{\frac{\left(n-1\right)}{\left(N-1\right)}-\frac{\left(n-2\right)\left(n-3\right)}{\left(N-2\right)\left(N-3\right)}\right\} + \frac{1}{144}\left(N^2-1\right)\left\{\frac{\left(N-3\right)-\left(n-2\right)\left(N+3\right)}{\left(N-1\right)^2} + \frac{\left(n-2\right)\left(n-3\right)\left(N^2-3\right)}{\left(N-1\right)^2\left(N-2\right)\left(N-3\right)}\right\}\right].$$

Similarly, substituting $m = m'$, $k_1 = m'k$ and $s = n/m'$ in Equation (B.7), one can get the following formula of variance of unbiased variance estimator with m' random starts for LSS in the presence of linear trend.

(6.3)
$$V\left(\hat{S}^2_{LSS}\right) = \frac{\beta^4\left(m'^2k^2-1\right)}{\left(N-1\right)^2}\frac{\left(m'N-n\right)^2 k}{m'^2\left(m'-1\right)} \times\left[\frac{\left(3m'^2k^2-7\right)}{240}\left\{\frac{\left(m'-1\right)}{\left(m'k-1\right)}-\frac{\left(m'-2\right)\left(m'-3\right)}{\left(m'k-2\right)\left(m'k-3\right)}\right\} + \frac{1}{144}\left(m'^2k^2-1\right)\left\{\frac{\left(m'k-3\right)-\left(m'-2\right)\left(m'k+3\right)}{\left(m'k-1\right)^2} + \frac{\left(m'-2\right)\left(m'-3\right)\left(m'^2k^2-3\right)}{\left(m'k-1\right)^2\left(m'k-2\right)\left(m'k-3\right)}\right\}\right].$$

6.1. Efficiency Comparison of $\hat{S}^2_{MSSM}$ and $\hat{S}^2_{SRS}$ in the Presence of Linear Trend

Due to complicated expressions given in Equation (B.7) and (6.2), theoretical comparison of $\hat{S}^2_{MSSM}$ and $\hat{S}^2_{SRS}$ is not easy. Therefore, a numerical comparison is carried out by considering the linear Model (3.1) and results are presented in Table 6.

Table 6: Efficiency of MSSM over SRS using linear model.

N	n	m	s	k_1	Efficiency	N	n	m	s	k_1	Efficiency
160	12	3	4	40	34.76	280	12	3	4	70	34.81
	14	7	2	80	6.72		15	3	5	56	65.24
	15	3	5	32	65.13		16	2	8	35	169.87
	18	9	2	80	6.98		18	9	2	140	6.98
	22	11	2	80	7.15		22	11	2	140	7.15
	24	3	8	20	250.30		24	3	8	35	250.80
	25	5	5	32	84.36		25	5	5	56	84.56
	26	13	2	80	7.27		26	13	2	140	7.27
	28	7	4	40	49.07		30	3	10	28	479.29
	30	3	10	16	478.57		32	4	8	35	299.77
	34	17	2	80	7.42		34	17	2	140	7.43
	35	7	5	32	94.01		36	9	4	70	52.04
	36	9	4	40	51.92		38	19	2	140	7.49
	38	19	2	80	7.48		42	3	14	20	1282.26
							44	11	4	70	53.96
							45	9	5	56	100.13
							46	23	2	140	7.57
							48	6	8	35	356.73
							49	7	7	40	252.94
							50	5	10	28	641.04

In Table 6, one can easily see that the lower the number of groups k_1, the higher is the efficiency, and vice versa. Note that different choices of α and β do not have any effect on the efficiencies as the parameters α and β will drop out from variance and efficiency expressions respectively.

7. CONCLUSION

The proposed MSSM design is based on adjusting the population units in groups. Thus, except the two extreme cases of this design, MSSM is neither completely systematic nor random but displaying the amalgamation of systematic and simple random sampling. In the two extreme cases, one of them becomes LSS and other SRS. The MSSM makes it possible to develop the modified expressions of all the results that relates to the LSS. A few such modifications are reported in Sections 2 and 3. A theoretical efficiency comparison of MSSM and SRS using the variances of mean in the presence of linear trend is carried out and is shown in Equation (3.1). This comparison clearly indicates that MSSM is more efficient than SRS.

In this study, population variance is unbiasedly estimated in MSSM for all possible combinations of N and n. An explicit expression for variance of unbiased variance estimator is also obtained in the proposed design. Moreover, it enables us to deduce the expressions for variance of unbiased variance estimator for LSS and SRS. Due to the complex nature of these expressions, theoretical comparison is not an easy task. Therefore, numerical comparison of MSSM and SRS is carried out in Sections 5 and 6. This numerical efficiency comparison is done for natural population, simulated population and linear model having a perfect linear trend. The results show that if populations (with linear or parabolic trend) are arranged with certain characteristics then MSSM is more efficient than SRS. However, in simulated populations, MSSM is almost equally efficient to SRS as units are not arranged in specific order. In this case, one can benefit from MSSM due to its simplicity and economical status. Furthermore, the findings reveal that the efficiency of MSSM is quite high for those combinations of N and n in which all population units are arranged in minimum number of groups.

APPENDIX A — Variance of $\hat{S}^2_{MSSM}$

The variance of $\hat{S}^2_{MSSM}$ can be written as

(A.1)

$$V\left(\hat{S}^2_{MSSM}\right) = \frac{1}{(N-1)^2}\left[\left(\frac{N}{m}\right)^2 V\Big(\sum_{u=1}^{m}\hat{\sigma}^2_{r_u}\Big) + \left(\frac{(N-s)}{m(m-1)}\right)^2 V\Big(\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}(\bar{y}_{r_u}-\bar{y}_{r_v})^2\Big) + 2\,\frac{N}{m}\frac{(N-s)}{m(m-1)}\,Cov\Big(\sum_{u=1}^{m}\hat{\sigma}^2_{r_u}, \sum_{u=1}^{m-1}\sum_{v=u+1}^{m}(\bar{y}_{r_u}-\bar{y}_{r_v})^2\Big)\right].$$

Note that

$$V\Big(\sum_{u=1}^{m}\hat{\sigma}^2_{r_u}\Big) = \sum_{u=1}^{m} V\left(\hat{\sigma}^2_{r_u}\right) + \sum_{u=1}^{m}\sum_{\substack{v=1\\ v\neq u}}^{m} Cov\left(\hat{\sigma}^2_{r_u}, \hat{\sigma}^2_{r_v}\right),$$

where

$$V(\hat{\sigma}^2_{r_u}) = \frac{1}{k_1}\sum_{u=1}^{k_1}\left(\hat{\sigma}^2_{r_u}-\bar{\sigma}^2\right)^2 = \frac{1}{k_1}\sum_{r=1}^{k_1}\left(\hat{\sigma}^2_{r}-\bar{\sigma}^2\right)^2 = \sigma_0^2 \qquad \text{(say)}$$

such that

$$\bar{\sigma}^2 = \frac{1}{k_1}\sum_{u=1}^{k_1}\hat{\sigma}^2_{r_u} = \frac{1}{k_1}\sum_{r=1}^{k_1}\hat{\sigma}^2_{r} \qquad \text{and} \qquad Cov\left(\hat{\sigma}^2_{r_u}, \hat{\sigma}^2_{r_v}\right) = -\frac{\sigma_0^2}{(k_1-1)}.$$

Thus

$$V\Big(\sum_{u=1}^{m}\hat{\sigma}^2_{r_u}\Big) = m\sigma_0^2\left(\frac{k_1-m}{k_1-1}\right). \tag{A.2}$$

Now consider

$$V\Big[\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}(\bar{y}_{r_u}-\bar{y}_{r_v})^2\Big] = \tag{A.3}$$

$$= \sum_{u=1}^{m-1}\sum_{v=u+1}^{m} V\left\{(\bar{y}_{r_u}-\bar{y}_{r_v})^2\right\} + 2\Bigg[\sum_{u=1}^{m}\sum_{\substack{v=1\\ v\neq u}}^{m}\sum_{\substack{u'=1\\ u'\neq u,v}}^{m} Cov\Big\{(\bar{y}_{r_u}-\bar{y}_{r_v})^2, (\bar{y}_{r_u}-\bar{y}_{r_{u'}})^2\Big\} + \sum_{u=1}^{m}\sum_{\substack{v=1\\ v\neq u}}^{m}\sum_{\substack{u'=1\\ u'\neq u,v}}^{m}\sum_{\substack{v'=1\\ v'\neq u,v,u'}}^{m} Cov\Big\{(\bar{y}_{r_u}-\bar{y}_{r_v})^2, (\bar{y}_{r_{u'}}-\bar{y}_{r_{v'}})^2\Big\}\Bigg],$$

where

$$V\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2 \;=\; \frac{2k_1}{(k_1-1)}\left\{\mu_4+\frac{k_1-3}{(k_1-1)}\mu_2^2\right\}, \tag{A.4}$$

such that

$$\mu_2=\frac{1}{k_1}\sum_{u=1}^{k_1}\left(\bar{y}_{r_u}-\mu\right)^2=\frac{1}{k_1}\sum_{r=1}^{k_1}\left(\bar{y}_r-\mu\right)^2 \qquad \text{and} \qquad \mu_4=\frac{1}{k_1}\sum_{r=1}^{k_1}\left(\bar{y}_r-\mu\right)^4.$$

$$Cov\left\{\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2,\left(\bar{y}_{r_u}-\bar{y}_{r_{u'}}\right)^2\right\} \;=\; \frac{k_1}{(k_1-1)}\left[\mu_4-\frac{k_1+3}{(k_1-1)}\mu_2^2\right]. \tag{A.5}$$

(A.6)

$$Cov\left\{\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2,\left(\bar{y}_{r_{u'}}-\bar{y}_{r_{v'}}\right)^2\right\} \;=\; \frac{-4k_1}{(k_1-2)(k_1-3)}\left[\mu_4-\frac{\left(k_1^2-3\right)}{(k_1-1)^2}\mu_2^2\right].$$

Putting (A.4), (A.5) and (A.6) in (A.3), we have

$$\begin{aligned}V\left[\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2\right] &= \binom{m}{2}\left[\frac{2k_1}{(k_1-1)}\left\{\mu_4+\frac{k_1-3}{(k_1-1)}\mu_2^2\right\}\right]\\ &\quad+\;2\left[m\binom{m-1}{2}\left\{\frac{k_1}{(k_1-1)}\left(\mu_4-\frac{k_1+3}{(k_1-1)}\mu_2^2\right)\right\}\right.\\ &\quad+\left\{\binom{\frac{m(m-1)}{2}}{2}-m\binom{m-1}{2}\right\}\\ &\quad\left.\times\left\{\frac{-4k_1}{(k_1-2)(k_1-3)}\left(\mu_4-\frac{\left(k_1^2-3\right)}{(k_1-1)^2}\mu_2^2\right)\right\}\right],\end{aligned}$$

or

$$\begin{aligned}V\left[\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2\right] &= \;m(m-1)k_1\left[\left\{\frac{(m-1)}{(k_1-1)}-\frac{(m-2)(m-3)}{(k_1-2)(k_1-3)}\right\}\mu_4\right.\\ &\quad+\left\{\frac{(k_1-3)-(m-2)(k_1+3)}{(k_1-1)^2}\right.\\ &\quad\left.\left.+\;\frac{(m-2)(m-3)\left(k_1^2-3\right)}{(k_1-1)^2(k_1-2)(k_1-3)}\right\}\mu_2^2\right].\end{aligned} \tag{A.7}$$

Also consider

$$\begin{aligned}Cov\left\{\sum_{u=1}^{m}\hat{\sigma}_{r_u}^2,\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2\right\} &=\\ &= \;E\left\{\sum_{u=1}^{m}\hat{\sigma}_{r_u}^2\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2\right\}\\ &\quad-E\left(\sum_{u=1}^{m}\hat{\sigma}_{r_u}^2\right)E\left\{\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2\right\},\end{aligned}$$

where

$$E\Big(\sum_{u=1}^{m}\hat{\sigma}_{r_u}^2\Big)=\sum_{u=1}^{m}E\left(\hat{\sigma}_{r_u}^2\right)=\,m\frac{1}{k_1}\sum_{u=1}^{k_1}\hat{\sigma}_{r_u}^2=\,m\frac{1}{k_1}\sum_{r=1}^{k_1}\hat{\sigma}_{r}^2\,=\,m\bar{\sigma}^2,$$

$$E\Big\{\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2\Big\}=\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}E\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2=\binom{m}{2}\frac{2k_1}{(k_1-1)}\,\mu_2$$

and

$$E\Big\{\sum_{u=1}^{m}\hat{\sigma}_{r_u}^2\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2\Big\}=\frac{m\,(m-1)}{(k_1-1)}\Big[\Big\{1+\frac{(m-2)\,(k_1-1)}{(k_1-2)}\Big\}k_1\bar{\sigma}^2\mu_2$$
$$+\Big\{1-\frac{(m-2)}{(k_1-2)}\Big\}\sum_{r=1}^{k_1}\hat{\sigma}_r^2\left(\bar{y}_r-\mu\right)^2\Big],$$

or

$$Cov\Big\{\sum_{u=1}^{m}\hat{\sigma}_{r_u}^2,\sum_{u=1}^{m-1}\sum_{v=u+1}^{m}\left(\bar{y}_{r_u}-\bar{y}_{r_v}\right)^2\Big\}=$$
$$=\frac{m\,(m-1)\,(k_1-m)}{(k_1-1)\,(k_1-2)}\Big\{\sum_{r=1}^{k_1}\hat{\sigma}_r^2\left(\bar{y}_r-\mu\right)^2-k_1\bar{\sigma}^2\mu_2\Big\}. \tag{A.8}$$

Putting (A.1), (A.7) and (A.8) in (A.1) and then simplifying, we have

(A.9)

$$V\left(\hat{S}_{MSSM}^2\right)=\frac{1}{m\,(N-1)^2}\Bigg[\frac{N^2\,(k_1-m)}{(k_1-1)}\,\sigma_0^2+\frac{(N-s)^2\,k_1}{(m-1)}$$
$$\times\Bigg[\Big\{\frac{(m-1)}{(k_1-1)}-\frac{(m-2)\,(m-3)}{(k_1-2)\,(k_1-3)}\Big\}\mu_4$$
$$+\Big\{\frac{(k_1-3)-(m-2)\,(k_1+3)}{(k_1-1)^2}+\frac{(m-2)\,(m-3)\left(k_1^2-3\right)}{(k_1-1)^2\,(k_1-2)\,(k_1-3)}\Big\}\mu_2^2\Bigg]$$
$$+\,2\frac{N\,(N-s)\,(k_1-m)}{(k_1-1)\,(k_1-2)}\Big\{\sum_{r=1}^{k_1}\hat{\sigma}_r^2\left(\bar{y}_r-\bar{Y}\right)^2-k_1\bar{\sigma}^2\mu_2\Big\}\Bigg].$$

APPENDIX B — Variance of $\hat{S}^2_{MSSM}$

Assuming the linear Model (3.1), the mean of the r^{th} $(r = 1, 2, ..., k_1)$ group can be written as

$$\bar{y}_r = \frac{1}{s}\sum_{i=1}^{s}\Big\{\alpha + \beta\big(r + (i-1)k_1\big)\Big\},$$

$$\bar{y}_r = \alpha + \beta\Big(r + \frac{1}{2}(s-1)k_1\Big), \tag{B.1}$$

$$\begin{aligned}\hat{\sigma}_r^2 &= \frac{1}{s}\sum_{i=1}^{s}\left\{\alpha + \beta\big(r + (i-1)k_1\big) - \alpha - \beta\Big(r + \frac{1}{2}(s-1)k_1\Big)\right\}^2 \\ &= \frac{1}{s}\sum_{i=1}^{s}\left\{\beta(i-1)k_1 - \beta\Big(\frac{1}{2}(s-1)k_1\Big)\right\}^2 \qquad \text{(B.2)} \\ &= \frac{1}{12}\beta^2 k_1^2(s^2-1),\end{aligned}$$

$$\bar{\sigma}_r^2 = \frac{1}{12}\beta^2 k_1^2(s^2-1), \tag{B.3}$$

$$\sigma_0^2 = 0, \tag{B.4}$$

$$\mu_2 = \frac{1}{k_1}\sum_{r=1}^{k_1}(\bar{y}_r - \mu)^2 = \frac{\beta^2}{12}\left(k_1^2 - 1\right) \tag{B.5}$$

and

$$\mu_4 = \frac{1}{k_1}\sum_{r=1}^{k_1}(\bar{y}_r - \mu)^4 = \beta^4\left(\frac{k_1^4}{80} - \frac{k_1^2}{24} + \frac{7}{240}\right), \tag{B.6}$$

where

$$\mu = \alpha + \beta\frac{N+1}{2}.$$

Putting Equations (B.1)–(B.6) in (A.9), we have

(B.7)

$$\begin{aligned}V\left(\hat{S}^2_{MSSM}\right) &= \frac{\beta^4\left(k_1^2-1\right)}{m\,(N-1)^2}\,\frac{(N-s)^2\,k_1}{(m-1)} \\ &\quad\times\left[\frac{\left(3k_1^2-7\right)}{240}\left\{\frac{(m-1)}{(k_1-1)} - \frac{(m-2)\,(m-3)}{(k_1-2)\,(k_1-3)}\right\} + \frac{1}{144}\left(k_1^2-1\right)\right. \\ &\quad\left.\times\left\{\frac{(k_1-3)-(m-2)\,(k_1+3)}{(k_1-1)^2}\,\frac{(m-2)\,(m-3)\left(k_1^2-3\right)}{(k_1-1)^2\,(k_1-2)\,(k_1-3)}\right\}\right].\end{aligned}$$

ACKNOWLEDGMENTS

The authors offer their sincere thanks to the two reviewers for their careful reading of the paper and their helpful suggestions.

REFERENCES

[1] Chang, H.J. and Huang, K.C. (2000). Remainder linear systematic sampling, *Sankhya*, **62**(B), 249–256.

[2] Gautschi, W. (1957). Some remarks on systematic sampling, *The Annals of Mathematical Statistics*, **28**(2), 385–394.

[3] Khan, Z.; Shabbir, J. and Gupta, S.N. (2013). A new sampling design for systematic sampling, *Communications in Statistics — Theory and Methods*, **42**(18), 3359–3370.

[4] Murthy, M.N. (1967). *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta, India.

[5] Naidoo, L.R.; North, D.; Zewotir, T. and Arnab, R. (2016). Multiple-start balanced modified systematic sampling in the presence of linear trend, *Communications in Statistics — Theory and Methods*, **45**(14), 4307–4324.

[6] Sampath, S. (2009). Finite population variance estimation under lss with multiple random starts, *Communication in Statistics — Theory and Methods*, **38**, 3596–3607.

[7] Sampath, S. and Ammani, S. (2012). Finite-population variance estimation under systematic sampling schemes with multiple random starts, *Journal of Statistical Computation and Simulation*, **82**(8), 1207–1221.

[8] Sethi, V.K. (1965). On optimum paring of units, *Sankhya*, **27**(B), 315–320.

[9] Singh, D.; Jindal, K.K. and Garg, J.N. (1968). On modified systematic sampling, *Biometrika*, **55**, 541–546.

[10] Subramani, J. and Gupta, S.N. (2014). Generalized modified linear systematic sampling scheme, *Hacettepe Journal of Mathematics and Statistics*, **43**(3), 529–542.

REVSTAT – Statistical Journal
Volume 16, Number 2, April 2018, 213–230

IMPROVING BAYESIAN MIXTURE MODELS FOR MULTIPLE IMPUTATION OF MISSING DATA USING FOCUSED CLUSTERING

Authors: Lan Wei
– In4mation Insights,
Needham, MA, USA

Jerome P. Reiter
– Department of Statistical Science, Duke University,
Box 90251, Durham, NC, 27708, USA
jerry@stat.duke.edu

Received: February 2017 Revised: September 2017 Accepted: October 2017

Abstract:

- We present a joint modeling approach for multiple imputation of missing continuous and categorical variables using Bayesian mixture models. The approach extends the idea of focused clustering, in which one separates variables into two sets before estimating the mixture model. Focus variables include variables with high rates of missingness and possibly other variables that could help improve the quality of the imputations. Non-focus variables include the remainder. In this way, one can use a rich sub-model for the focus set and a simpler model for the non-focus set, thereby concentrating fitting power on the variables with the highest rates of missingness. We present a procedure for specifying which variables with low rates of missingness to include in the focus set. We examine the performance of the imputation procedure using simulation studies based on artificial data and on data from the American Community Survey.

Key-Words:

- *incomplete; nonparametric; nonresponse; survey; tensor.*

1. INTRODUCTION

Nonparametric Bayesian (NB) mixture models are useful tools for analyzing complicated data ([13], [5], [14], [3], [2]). They are especially useful as engines for multiple imputation (MI, [16], [11], [18], [9], [12], [10], [7]). NB mixture models are flexible enough to capture complex relationships among the variables, which is advantageous in MI contexts where one seeks to create completed datasets for use in multiple analyses.

In many contexts, only a few variables have high rates of missingness, and other variables are nearly or completely observed. This can create estimation difficulties when using mixture models as MI engines. In particular, with modest sample sizes and many variables, mixture models have the potential to fit the distribution of some variables well at the expense of others ([6], [19], [4]). The mixture model easily could expend its fitting power on the marginal distribution of the (nearly) completely observed variables at the expense of the distribution of the variables with high rates of missingness ([4],[20]), which could lead to poor quality imputations.

To get around this, [4] suggest using mixture models with focused clustering. Using the nomenclature in [4], the variables with high rates of missing data are called focus variables, and the others are called remainder variables. In focused clustering, the mixture model includes one set of cluster indicators for focus variables and a second set for remainder variables. The two sets are connected using a tensor factorization prior ([15]). In this way, one can use a rich sub-model for the focus set and a simpler model for the remainder set, thereby concentrating fitting power on the variables with the highest rates of missingness.

In this article, we enhance the focused clustering approach for MI to facilitate higher quality imputations. In particular, we expand the definition of focus variables to include variables with high fractions of missing data and (nearly) completely observed variables that could improve the quality of the imputations for the variables with high rates of missingness; we label the resulting set with $\mathcal{F}$. We define the non-focus variables to include those not in $\mathcal{F}$; we label these as $\mathcal{NF}$. We specify the variables to include in $\mathcal{F}$ as follows. First, we automatically put all variables with high fractions of missing values in $\mathcal{F}$. For each variable not automatically in $\mathcal{F}$, we compute its mutual information with the variables automatically in $\mathcal{F}$. We move variables with high mutual information values into $\mathcal{F}$; the remaining variables we put in $\mathcal{NF}$. We make these decisions in one step, including all variables with high mutual information values in $\mathcal{F}$. We refer to this strategy as *M*ove. We use *Stay* to refer to the strategy of putting only variables with high fractions of missingness in $\mathcal{F}$. Because *Move* allows local dependence among the variables with high amounts of missing values and (nearly) completely

observed variables that can be used to predict the missing values, it can improve accuracy and, in some cases, computational efficiency.

The remainder of this article is organized as follows. In Section 2, we present the focused clustering model, which we abbreviate as HCMM-FNF for hierarchically coupled mixture model with focus/non-focus variables, and motivate the potential benefits of *Move*. In Section 3, we illustrate when *Move* engenders benefits using four simple simulation scenarios. In Section 4, we apply the strategies to data sampled from the American Community Survey. In Section 5, we conclude with a brief summary of findings.

2. SPECIFICATION OF HCMM-FNF

We indicate continuous variables with Y and categorical variables with X. We use a superscript F to denote focus variables and the superscript NF to denote non-focus variables. Thus, $Y^{(F)}$, $X^{(F)}$, $Y^{(NF)}$ and $X^{(NF)}$ are the focus continuous, focus categorical, non-focus continuous, and non-focus categorical variables, respectively. For purposes of explaining HCMM-FNF, here we assume that $\mathcal{F}$ and $\mathcal{NF}$ have been pre-specified.

For each observation $i = 1, ..., n$, we have $Y_i^{(F)} = \big(Y_{i1}^{(F)}, ..., Y_{iq^{(F)}}^{(F)}\big)^T$, $X_i^{(F)} = \big(X_{i1}^{(F)}, ..., X_{ip^{(F)}}^{(F)}\big)^T$, $Y_i^{(NF)} = \big(Y_{i1}^{(NF)}, ..., Y_{iq^{(NF)}}^{(NF)}\big)^T$, and $X_i^{(NF)} = \big(X_{i1}^{(NF)}, ..., X_{ip^{(NF)}}^{(NF)}\big)^T$. Let D_i be a regression design matrix containing the main effects of $X_i^{(F)}$, $Y_i^{(NF)}$, and $X_i^{(NF)}$. A similar regression approach is proposed by [15]. HCMM-FNF can be described as follows.

$$\big(Y_i^{(F)} | D_i, H_i^{(FY)} = a, -\big) \sim \mathcal{N}\big(y_i^{(F)} | D_i B_a^{(F)}, \Sigma_a^{(F)}\big), \tag{2.1}$$

$$Pr\big(X_i^{(F)} = x_i^{(F)} \,|\, H_i^{(FX)} = b, -\big) = \prod_{j=1}^{p^{(F)}} \psi_{b, x_{ij}^{(F)}}^{(F)(j)}, \tag{2.2}$$

$$\big(Y_i^{(NF)} | H_i^{(NF)} = h, -\big) \sim \mathcal{N}\big(y_i^{(NF)} | B_h^{(NF)}, \Sigma_h^{(NF)}\big), \tag{2.3}$$

$$Pr\big(X_i^{(NF)} = c_i^{(NF)} \,|\, H_i^{(NF)} = h, -\big) \sim \prod_{j=1}^{p^{(NF)}} \psi_{h, x_{ij}^{(NF)}}^{(NF)(j)}, \tag{2.4}$$

$$Pr\big(H_i^{(FY)} = a,\, H_i^{(FX)} = b \,|\, Z_i = z\big) = \phi_{z,a}^{(FY)} \phi_{z,b}^{(FX)}, \tag{2.5}$$

$$Pr\big(H_i^{(NF)} = h \,|\, Z_i = z\big) = \phi_{z,h}^{(NF)}, \tag{2.6}$$

$$Pr\big(Z_i = z\big) = \lambda_z. \tag{2.7}$$

$H_i^{(FY)} \in \{1, ..., k^{(FY)}\}$ is the mixture component index of $Y_i^{(F)}$. $H^{(FX)} \in \{1, ..., k^{(FX)}\}$ is the mixture component index of $X_i^{(F)}$. $H_i^{(NF)} \in \{1, ..., k^{(NF)}\}$

is the mixture component index of $Y_i^{(NF)}$ and $X_i^{(NF)}$. $Z_i \in \{1, ..., k^{(Z)}\}$ is the mixture component index of $H_i^{(F)}$ and $H_i^{(NF)}$. $B_a^{(F)}$ and $\Sigma_a^{(F)}$ are the matrix of regression coefficients and the covariance matrix in $H_i^{(FY)} = a$. $\psi_{b,x_{ij}^{(F)}}^{(F)(j)}$ is the probability of $X_{ij}^{(F)} = x_{ij}^{(F)}$ in $H_i^{(FX)} = b$. $B_h^{(NF)}$ and $\Sigma_h^{(NF)}$ are the mean vector and the covariance matrix in $H_i^{(NF)} = h$. Here, $\Sigma_h^{(NF)}$ is a diagonal matrix with non-zero entries $\big(\eta_{h,1}^{(NF)}, ..., \eta_{h,q^{(NF)}}^{(NF)}\big)$. Thus, the variables in $Y_i^{(NF)}$ are conditionally independent. Finally, $\psi_{h,x_{ij}^{(NF)}}^{(NF)(j)}$ is the probability of $X_{ij}^{(NF)} = x_{ij}^{(NF)}$ in $H_i^{(FX)} = h$.

To allow closed-form expressions for the posteriors, we take conjugacy into consideration when specifying the prior distributions. For the multinomial variables, we have

$$\psi_b^{(F)(j)} \overset{i.i.d.}{\sim} Dir\big(\gamma_{b,1}^{(j)}, ..., \gamma_{b,d_j^{(F)}}^{(j)}\big), \tag{2.8}$$

$$\psi_h^{(NF)(j)} \overset{i.i.d.}{\sim} Dir\big(\gamma_{h,1}^{(j)}, ..., \gamma_{h,d_j^{(NF)}}^{(j)}\big) \tag{2.9}$$

$$\big(\gamma_{b,1}^{(j)}, ..., \gamma_{b,d_j^{(F)}}^{(j)}\big)^T = \big(1/d_j^{(F)}, ..., 1/d_j^{(F)}\big)^T, \tag{2.10}$$

$$\big(\gamma_{h,1}^{(j)}, ..., \gamma_{h,d_j^{(NF)}}^{(j)}\big)^T = \big(1/d_j^{(NF)}, ..., 1/d_j^{(NF)}\big)^T, \tag{2.11}$$

For the multivariate normal variables, we have

$$Pr\big(B_a^{(F)}, \Sigma_a^{(F)}\big) = \mathcal{N}\big(B_0^{(F)}, I, T_B^{(F)}\big) \times \mathcal{IW}\big(\nu^{(F)}, \Sigma^{(F)}\big), \tag{2.12}$$

$$Pr\big(B_h^{(NF)}\big) = \mathcal{N}\big(B_0^{(NF)}, T_B^{(NF)}\big), \tag{2.13}$$

$$Pr\big(\eta_{h,j}^{(NF)}\big) = \mathcal{IG}\big(\nu^{(NF)}, \eta_j^{(NF)}\big), \tag{2.14}$$

where $T_B^{(F)} = Diag\big(\tau_1^{(F)}, ..., \tau_{q^{(F)}}^{(F)}\big)$ and $T_B^{(NF)} = Diag\big(\tau_1^{(NF)}, ..., \tau_{q^{(NF)}}^{(NF)}\big)$, and

$$\tau_j^{(F)} \overset{i.i.d.}{\sim} \mathcal{G}\big(\alpha_{\tau^{(F)}}, \beta_{\tau^{(F)}}\big), \tag{2.15}$$

$$\tau_j^{(NF)} \overset{i.i.d.}{\sim} \mathcal{G}\big(\alpha_{\tau^{(NF)}}, \beta_{\tau^{(NF)}}\big). \tag{2.16}$$

For the hyper-prior distributions, we have

$$\big(B_0^{(F)}, \Sigma^{(F)}\big) \sim \mathcal{N}\big(0, I, {\sigma_0^{(F)}}^2 I\big) \times \mathcal{W}\big(\omega^{(F)}, \Sigma_0^{(F)}\big), \tag{2.17}$$

$$\big(B_0^{(NF)}\big) \sim \mathcal{N}\big(0, {\sigma_0^{(NF)}}^2 I\big), \tag{2.18}$$

$$\big(\eta_j^{(NF)}\big) \sim \mathcal{IG}\big(\nu^{(NF)}, \eta_0^{(NF)}\big). \tag{2.19}$$

We let $\nu^{(F)} = q^{(F)} + 2$, $\nu^{(NF)} = 2$, $\omega^{(F)} = q^{(F)} + 1$, $\omega^{(NF)} = 1$, $\Sigma_0^{(F)} = I/(q^{(F)} + 1)$, and $\eta_0^{(NF)} = 1$.

The hierarchical priors for the latent variables follow a truncated version of the stick-breaking construction of the Dirichlet process ([17], [8]). We have

$$\phi_{z,a}^{(FY)} = V_{z,a}^{(FY)} \prod_{l<a} (1 - V_{z,l}^{(FY)}), \quad V_{z,a}^{(FY)} \overset{i.i.d.}{\sim} \mathcal{B}(1, \beta^{(FY)}), \quad V_{z,k^{(FY)}}^{(FY)} = 1, \tag{2.20}$$

$$\phi_{z,b}^{(FX)} = V_{z,b}^{(FX)} \prod_{l<b} (1 - V_{z,l}^{(FX)}), \quad V_{z,b}^{(FX)} \overset{i.i.d.}{\sim} \mathcal{B}(1, \beta^{(FX)}), \quad V_{z,k^{(FX)}}^{(FX)} = 1, \tag{2.21}$$

$$\phi_{z,h}^{(NF)} = V_{z,h}^{(NF)} \prod_{l<h} (1 - V_{z,l}^{(NF)}), \quad V_{z,h}^{(NF)} \overset{i.i.d.}{\sim} \mathcal{B}(1, \beta^{(NF)}), \quad V_{z,k^{(NF)}}^{(NF)} = 1, \tag{2.22}$$

$$\lambda_z = W_z \prod_{l<z} (1 - W_l), \quad W_z \overset{i.i.d.}{\sim} \mathcal{B}(1, \alpha), \quad W_{k^{(Z)}} = 1. \tag{2.23}$$

Details about the method of fitting the model can be found in Chapter 4 of [20].

Figure 1 is a graphical representation of HCMM-FNF. It is apparent that dependence between $X^{(F)}$ and all variables in $\mathcal{NF}$ is captured only by the lowest level of mixture components, which could make accurate estimation of these associations difficult. Dependence between $Y^{(F)}$ and all variables in $\mathcal{NF}$ is captured via the component regressions and the lowest level of mixture components.

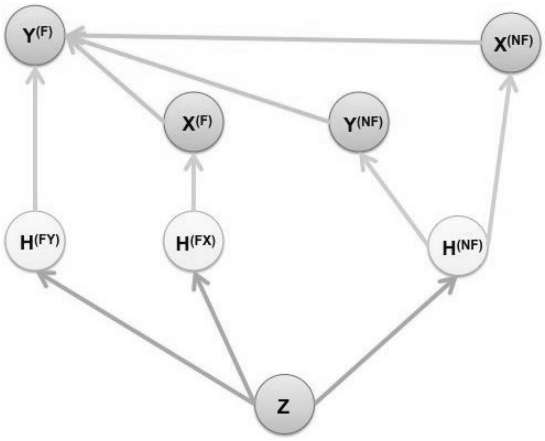


Figure 1: Graphical model representation of HCMM-FNF. $X^{(F)}$, $Y^{(F)}$, $X^{(NF)}$, and $Y^{(NF)}$ are the observed categorical and continuous variables. $H^{(F)}$ and $H^{(NF)}$ are the mixture components of $\mathcal{F}$ and $\mathcal{NF}$ variables, respectively. Z is the mixture component for $H^{(F)}$ and $H^{(NF)}$.

While this encodes dependence between $Y^{(F)}$ and all variables in $\mathcal{NF}$, we expect HCMM-FNF to do a better job capturing the joint distribution among variables within $\mathcal{F}$ than the relationships of $Y^{(F)}$ with variables in $\mathcal{NF}$, as the variables within $\mathcal{F}$ share mixture components directly. This suggests that when the associations between some variables in $Y^{(F)}$ and $Y^{(NF)}$ are strong or nonlinear, it may be advantageous to put all those variables in $\mathcal{F}$. Similarly, when $Y^{(F)}$ and

$X^{(NF)}$ are highly associated, moving $X^{(NF)}$ to $\mathcal{F}$ may improve the estimation of the associations between $Y^{(F)}$ and $X^{(NF)}$. Similarly, when some variables in $Y^{(NF)}$ are highly associated with $X^{(F)}$, or when some variables in $X^{(NF)}$ are highly associated with $X^{(F)}$, moving them to $\mathcal{F}$ could help the model estimate the associations.

These observations motivate why *Move* could lead to improved estimation over *Stay*. We now explore that possibility using simulation studies.

3. SIMULATION STUDIES

We investigate the potential of *Move* to improve the quality of imputations using four simple scenarios. To describe each scenario, let (F_0) index the focus variables automatically included in $\mathcal{F}$, i.e., those with high rates of missing values, and (NF_0) index the other variables. The sets of variables defined by (F_0) and (NF_0), which we call $\mathcal{F}_0$ and $\mathcal{NF}_0$, respectively, are those used in *Stay*. In *Move*, we put some variables in $\mathcal{NF}_0$ in $\mathcal{F}$.

3.1. Simulation scenarios and evaluation metrics

In Scenario 1, we make variables in $X^{(NF_0)}$ highly associated with some variables in $X^{(F_0)}$. We generate six binary $X^{(NF_0)}$ variables from an arbitrarily chosen joint distribution, constructed from a mixture of products of multinomial distributions. To create the dependencies between the categorical variables in $\mathcal{F}_0$ and $\mathcal{NF}_0$, we generate four $X^{(F_0)}$ variables according to Bernoulli distributions with $Pr\big(X_j^{(F_0)} = x \,|\, X_j^{(NF_0)} = x\big) = 0.9$, with $x \in \{1, 2\}$ for $j = 1, ..., 4$. Under *Move*, we put $\big(X_1^{(NF_0)}, ..., X_4^{(NF_0)}\big)$ in $\mathcal{F}$.

In Scenario 2, we make some variables in $Y^{(NF_0)}$ highly associated with variables in $X^{(F_0)}$. We generate six $Y^{(NF_0)}$ variables from an arbitrary mixture of normal distributions. We create four binary $X^{(F_0)}$ variables from Bernoulli distributions with

$$log\left(\frac{Pr\big(X_j^{(F_0)} = 2 \,|\, Y_j^{(NF_0)} = y_j^{(NF_0)}\big)}{Pr\big(X_j^{(F_0)} = 1 \,|\, Y_j^{(NF_0)} = y_j^{(NF_0)}\big)}\right) = y_j^{(NF_0)}, \tag{3.1}$$

for $j = 1, ..., 4$. Under *Move*, we put $\big(Y_1^{(NF_0)}, ..., Y_4^{(NF_0)}\big)$ in $\mathcal{F}$.

In Scenario 3, we make some variables in $X^{(NF_0)}$ highly associated with $Y^{(F_0)}$. We generate six binary $X^{(NF_0)}$ variables from an arbitrarily chosen mixture of products of multinomial distributions. We generate four $Y^{(F_0)}$ according to

$\big(Y_j^{(F_0)}|X_j^{(NF_0)}= x_j^{(NF_0)}\big) \sim \mathcal{N}\big(y_j^{(F_0)}|x_j^{(NF_0)}, 0.005\big)$, with $j = 1, ..., 4$. Under *Move*, we put $\big(X_1^{(NF_0)}, ..., X_4^{(NF_0)}\big)$ in $\mathcal{F}$.

In Scenario 4, we make some variables in $Y^{(NF_0)}$ highly associated with $Y^{(F_0)}$. We generate six $Y^{(NF_0)}$ variables from an arbitrarily chosen mixture of normal distributions. We generate four $Y^{(F_0)}$ according to $\big(Y_j^{(F_0)}|Y_j^{(NF_0)}= y_j^{(NF_0)}\big) \sim \mathcal{N}\big(0.9y_j^{(NF_0)}, 0.005\big)$, for $j = 1, ..., 4$. Under *Move*, we put $\big(Y_1^{(NF_0)}, ..., Y_4^{(NF_0)}\big)$ in $\mathcal{F}$.

We use two evaluation metrics in the simulations. Let $q_{k,j,l}^{(s)}$ be the k^{th} quantity of interest in the j^{th} repeated sample for the l^{th} imputation. The superscript (s) indicates that the estimate is from *Stay*. Similarly, we define $q_{k,j,l}^{(m)}$ for the estimate obtained from *Move*. Notations without any superscripts and subscript l, such as $q_{k,j}$, stand for the quantities from the truth, defined as the complete data without any missing values.

Metric I: We define the absolute differences as $d_{k,j,l}^{(s)} = |q_{k,j,l}^{(s)} - q_{k,j}|$ for *Stay* and $d_{k,j,l}^{(m)} = |q_{k,j,l}^{(m)} - q_{k,j}|$ for *Move*. We compute $d_{k,j}^{(s)} = (1/L)\sum_{l=1}^{L} d_{k,j,l}^{(s)}$ and $d_{k,j}^{(m)} = (1/L)\sum_{l=1}^{L} d_{k,j,l}^{(m)}$. For each quantity, we conduct a paired t-test of the hypothesis $H_0 : \mu_k^{(s)} = \mu_k^{(m)}$, where $\mu_k^{(s)}$ is the population mean of $d_{k,j}^{(s)}$ and $\mu_k^{(m)}$ is the population mean of $d_{k,j}^{(m)}$. When the p-value is below 0.01, we consider the difference between *Stay* and *Move* statistically significant.

Metric II: We define the percentage changes as $\Delta d_{k,j,l}^{(s)} = \frac{q_{k,j,l}^{(s)} - q_{k,j}}{q_{k,j}} \times 100\%$ for *Stay* and $\Delta d_{k,j,l}^{(m)} = \frac{q_{k,j,l}^{(m)} - q_{k,j}}{q_{k,j}} \times 100\%$ for *Move*. This metric is useful when the quantities of interest are not in the same units. For each quantity k, we let $\Delta d_k^{(s)} = (1/JL)\sum_{j=1}^{J}\sum_{l=1}^{L} \Delta d_{k,j,l}^{(s)}$ and $\Delta d_k^{(m)} = (1/JL)\sum_{j=1}^{J}\sum_{l=1}^{L} \Delta d_{k,j,l}^{(m)}$. We then draw box plots for all $\{\Delta d_k^{(s)}\}$ and $\{\Delta d_k^{(m)}\}$ of the same type. For example, we draw box plots of $\{\Delta d_k^{(s)}\}$ and $\{\Delta d_k^{(m)}\}$ for all possible correlations between $Y^{(F)}$ and $Y^{(NF)}$.

3.2. Results

For each scenario, we generate 100 independent datasets comprising $n = 1,000$ observations. For some variables, we make 50% of values missing completely at random (MCAR) and automatically put them in $\mathcal{F}_0$; for the remainder, we make only 1% MCAR and put them in $\mathcal{NF}_0$. In each incomplete dataset, we fit HCMM-FNF with *Move* and *Stay*, using 25,000 iterations as burn-in, which is sufficient based on standard diagnosis of MCMC convergence. After burnin, we run the chains for 1,000 iterations, and from these keep $L = 10$ imputations spaced 100 iterations apart.

Figure 2 displays results from Scenario 1 for bivariate probabilities between the categorical variables in $\mathcal{F}_0$ and $\mathcal{NF}_0$. Generally, the cell probabilities are estimated more accurately under *Move* than *Stay*. The improvements are most noticeable in the probabilities involving $\left(X_j^{(NF_0)}, X_j^{(F_0)}\right)$ where $j = 1, ..., 4$. Detailed investigation of the box plots for small values of Metric II indicates that the percentage changes under *Move* are generally smaller than those under *Stay*.

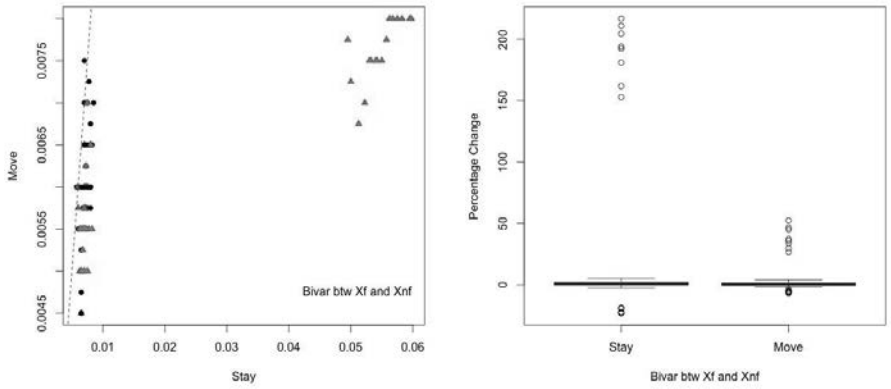


Figure 2: Bivariate cell probabilities for *Stay* and *Move* in Scenario 1. The left plot shows Metric I, where triangles correspond to p-values below 0.01 when testing for average differences in the two strategies. The right plot shows Metric II. The median of the relative differences is 0.0 for both *Stay* and *Move*.

In Scenario 2, we examine the coefficients of the logistic regressions of each $X^{(F_0)}$ variable on each $Y^{(NF_0)}$ variable. As evident in Figure 3, these coefficients are estimated more accurately in *Move* than in *Stay*. The accuracy gains are largest for the coefficients involving $\left(X_j^{(F_0)}, Y_j^{(NF_0)}\right)$ where $j = 1, ..., 4$.

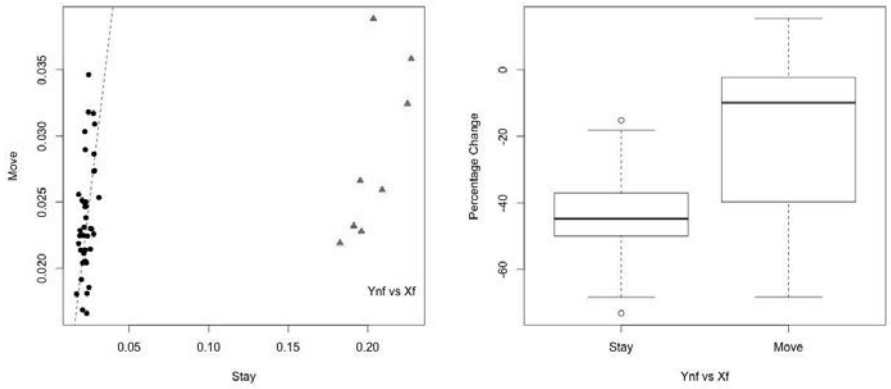


Figure 3: Coefficients in logistic regressions for *Stay* and *Move* in Scenario 2. The left plot shows Metric I, where triangles correspond to p-values below 0.01 when testing for average differences in the two strategies. The right plot shows Metric II. The median of the relative differences is -44.8 for *Stay* and -9.9 for *Move*.

In Scenario 3, we are interested in the associations between the variables in $Y^{(F_0)}$ and $X^{(NF_0)}$. We measure these associations using logistic regressions of $X_j^{(NF_0)}$ on $Y_k^{(F_0)}$ for $j \in \{1, ..., 4\}$ and $k \in \{1, ..., 6\}$. As evident in Figure 4, there are no significant differences between *Move* and *Stay* on Metric I. The box plots for Metric II show that the two medians are close, although the spread of values for *Move* is smaller than that for *Stay*.

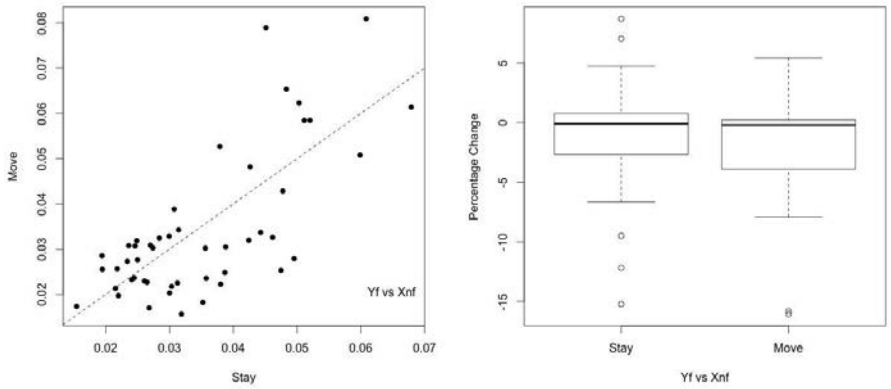


Figure 4: Coefficients in logistic regressions for *Stay* and *Move* in Scenario 3. The left plot shows Metric I, and the right plot shows Metric II. The median of the relative differences is −0.09 for *Stay* and −0.10 for *Move*.

For Scenario 4, Figure 5 displays results for the pairwise correlations of variables in $Y^{(F_0)}$ and $Y^{(NF_0)}$. There are no significant differences between *Move* and *Stay* for Metric I or Metric II.

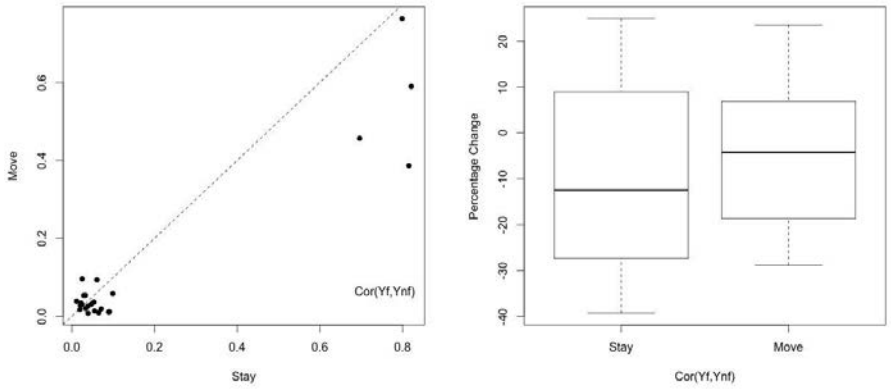


Figure 5: Pairwise correlations for *Move* and *Stay* in Scenario 4. The left plot shows Metric I, and the right plot shows Metric II. The median of the relative differences is −12.4 for *Stay* and −4.3 for *Move*.

3.3. Summary of results

When using *Stay*, associations between $X^{(F_0)}$ and $X^{(NF_0)}$ are estimated only through the tensor factorization. Apparently, in Scenario 1 this is not sufficient to capture the dependence. In contrast, by using common mixture components for all the categorical variables in $\mathcal{F}$, *Move* captures the dependence structure in Scenario 1 more effectively than *Stay*. We reach similar findings for Scenario 2, in which the local dependence enabled by *Move* captures associations involving $X^{(F_0)}$ and $Y^{(NF_0)}$ more effectively than relying only on the tensor factorization to capture the dependence. These results are in accord with the motivation we gave at the end of Section 2 for moving some (nearly) completely observed variables to $\mathcal{F}$.

For the associations between $Y^{(F_0)}$ and $\mathcal{NF}_0$, *Move* does not offer significant benefits over *Stay* in Scenarios 3 and 4. Apparently, *Stay* adequately incorporates the dependence between $Y^{(F_0)}$ and $\left(X^{(F_0)}, X^{(NF_0)}, Y^{(NF_0)}\right)$ through the mixture component regressions, so that moving variables to $\mathcal{F}$ does not noticeably improve the imputation quality. We also tried four modifications of these scenarios that use nonlinear associations between $Y^{(F_0)}$ and variables in $\mathcal{NF}_0$; see [20] for details of the designs. The performances of *Move* and *Stay* were qualitatively similar. Apparently, by using mixture distributions for the focus variables, we potentially can capture nonlinear relationships among the continuous focus variables.

4. EMPIRICAL STUDY

The findings in Section 3.3 are based on stylized simulation scenarios designed to clarify when *Move* can be advantageous. Further, in the studies we moved the nearly completely observed variables known to have strong associations with the variables in $\mathcal{F}_0$; in genuine settings we need empirical measures to identify these variables. In this section we present such measures and investigate whether or not similar behavior holds for genuine data.

4.1. Illustrative Data: The American Community Survey

The American Community Survey (ACS), an ongoing survey conducted by the U.S. Census Bureau, collects demographic, housing, social, and economic data from sampled households along with information on the people who live in these households. It is a rich and dynamic resource for public policy decision making

and analysis. Researchers can access public use files from the Integrated Public Use Microdata Series (IPUMS, usa.ipums.org). Relationships among variables in the ACS can be complex and difficult to capture with standard imputation models ([15]). Thus, we can benefit from using HCMM-FNF for imputation modeling.

We subset the ACS data to include only household heads who own their living units, were employed during the year of 2010 in the state of North Carolina, and have complete data; this subset has $19,492$ cases. We systematically sample $1,026$ household heads as our working dataset. To facilitate reasonable computation time, we choose the 16 variables in Table 1. Since IPUMS processes the raw data, the percentage of missing values for each variable in the IPUMS file is less than 2%. We therefore introduce additional missing values for purposes of the empirical study.

Before presenting results, we note that we repeated both studies on a second random sample of $1,026$ qualifying household heads. The patterns are very similar to the ones presented here; see Chapter 4 of [20] for details.

Table 1: Variables in ACS empirical study. First four variables are for households; the remainder are for the head of the household. *Cts* is short for continuous, and *Cat* is short for categorical. # Levels is the number of levels of the categorical variable. PROPTX99 is categorical with a large number of levels, and is modeled as such. It is treated as continuous when we report results.

Name	Label	Cts./Cat.[#Levels]
PROPTX99	Annual property taxes	Cat[67]
COSTELEC	Annual electricity cost	Cts
COSTGAS	Annual gas cost	Cts
COSTWATR	Annual water cost	Cts
AGE	Age	Cts
SEX	Sex	Cat[2]
MARST	Marital status	Cat[6]
RACE	Race	Cat[7]
HCOVANY	Any health insurance coverage	Cat[2]
EDUC	Educational attainment	Cat[9]
SCHLTYPE	Public or private school	Cat[3]
INCTOT	Total personal income	Cts
OCCSCORE	Occupational income score	Cts
PWTYPE	Place of work: metropolitan status	Cat[5]
MIGRATE1	Migration status, 1 year	Cat[4]
DIFFSENS	Vision or hearing difficulty	Cat[2]

4.2. Studies

As the measure to determine which variables to move into $\mathcal{F}$, we use the relative mutual information. For any two continuous variables A and B, the mutual information is

$$I(A,B) = \int_B \int_A p(a,b)\ log\left(\frac{p(a,b)}{p(a)p(b)}\right) da\, db. \tag{4.1}$$

The relative mutual information with respect to a variable A is a ratio of $I(A,B)$ over $I(A,A)$. For categorical variables, we replace the integrals with summations.

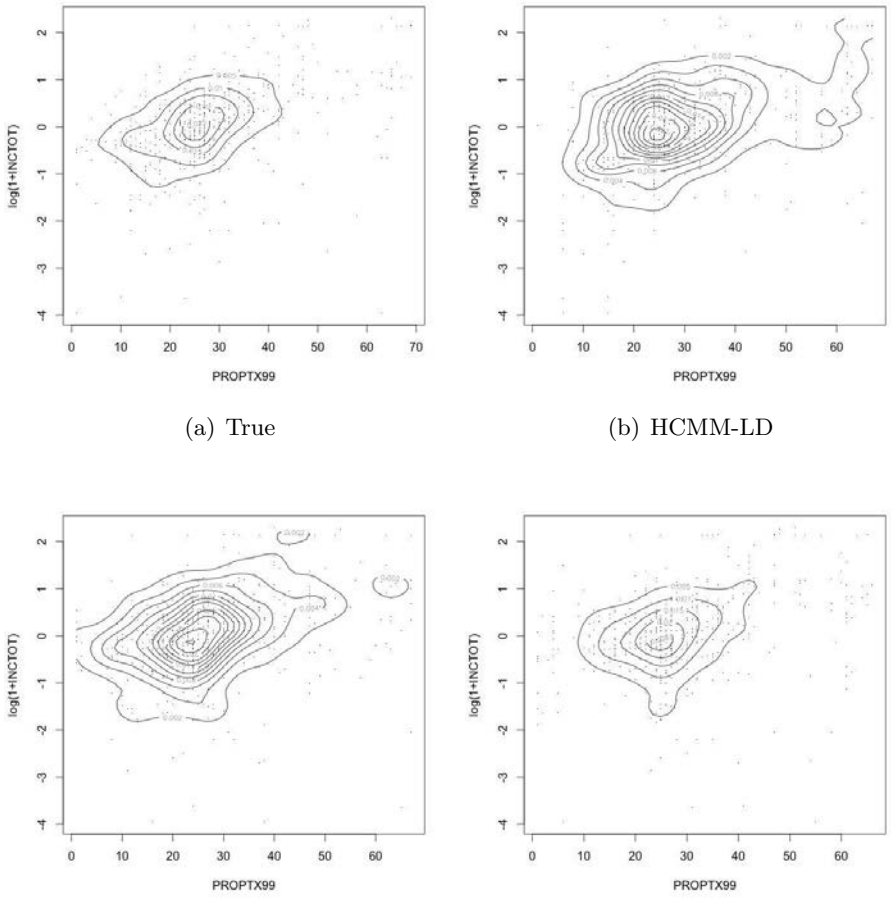
We run two studies, which we call the high and low mutual information studies. In each study, we impute the missingness in the working dataset using three models: HCMM-FNF with *Stay*, HCMM-FNF with *Move*, and the mixture model of [15], which we label HCMM-LD. HCMM-LD does not use any focused clustering, essentially putting all variables in $\mathcal{F}$. We use the performance of HCMM-LD as a benchmark for *Stay* and *Move*.

High Mutual Information (HMI) Study

We begin with a study in which variables in $\mathcal{NF}_0$ are predictive of variables in $X^{(F_0)}$, i.e., they share high amounts of mutual information. From the categorical variables in Table 1, we assign EDUC and PROPTX99 to have 50% values MCAR and thus to be in $\mathcal{F}_I$ automatically. We assign INCTOT, OCCSCORE, AGE, COSTELEC, COSTGAS, and COSTWATR as $Y^{(NF_0)}$, and the remaining variables as $X^{(NF_0)}$. Variables in $\mathcal{NF}_0$ have 1% values MCAR.

INCTOT and OCCSCORE have relatively high mutual information with EDUC and PROPTX99 with values at 0.26 and 0.22, respectively. All other values are 0.11 or lower, with all but two being below 0.05. Thus, we add INCTOT and OCCSCORE to the focus variables under *Move*. We analyze the marginal probabilities of PROPTX99 and EDUC, and pay special attention to associations between the variables in $\mathcal{F}$ after *Move*.

Figure 6 displays contour plots from the kernel density estimates of the standardized values of $log(1+INCTOT)$ and PROPTX99 for the missing observations. The true density is unimodal, concentrated in the area with PROPTX99 from $(5, 45)$ and $log(1+INCTOT)$ from $(-1.5, 1.2)$. By comparison, the completed data density estimates under HCMM-LD and *Stay* have a large spread and distorted contours. The density estimate under *Move* looks most similar to the truth.

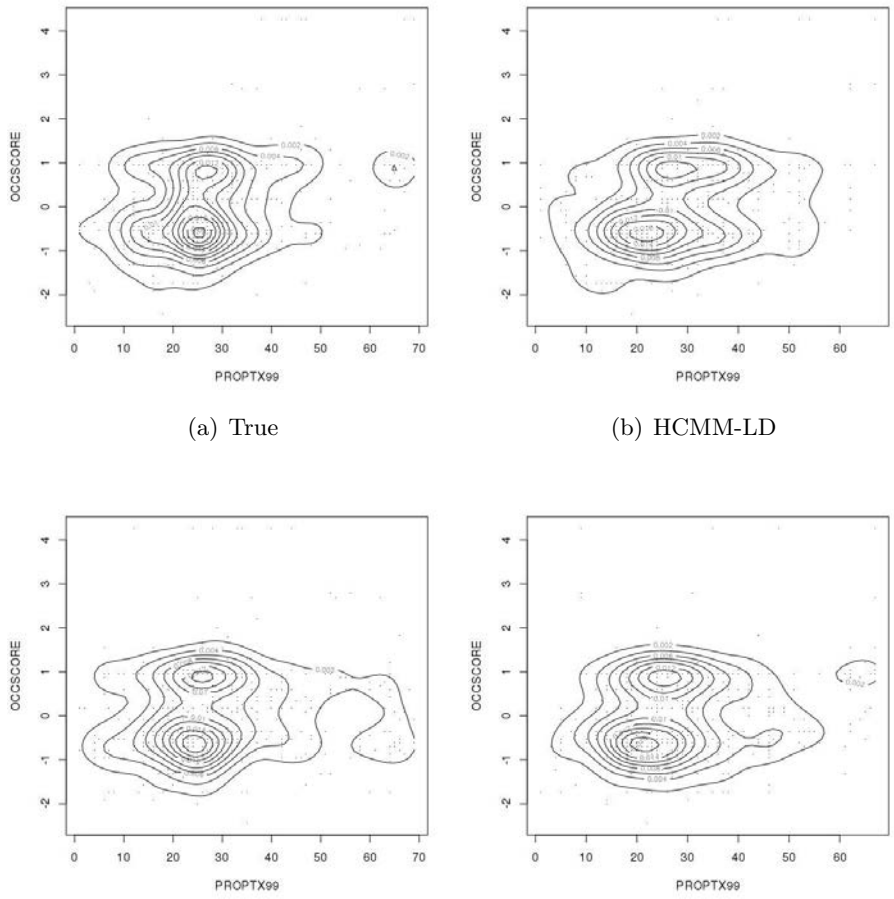


(a) True (b) HCMM-LD

(c) *Stay* (d) *Move*

Figure 6: Contour plots from the kernel density estimates of $log(1 + INCTOT)$ (standardized) and PROPTX99 for the missing observations in the *HMI* study. Each completed-data plot is from one randomly selected dataset.

Figure 7 displays the kernel density estimate of the standardized OCCSCORE and PROPTX99 for the missing observations. The true density has two high density, connected modes and one low density, isolated mode. The small mode reflects household heads whose occupational score is around 1 (41 on the original scale) and pay a high amount for their property taxes. Both HCMM-LD and *Stay* have trouble capturing this isolated mode; *Move* captures it more effectively than the other models. There are no significant differences among the three models for other quantities, including the marginal cell counts of EDUC and the bivariate associations involving EDUC. Details can be found in Chapter 4 of [20].



(a) True (b) HCMM-LD

(c) *Stay* (d) *Move*

Figure 7: Contour plots from the kernel density of OCCSCORE (standardized) and PROPTX99 for the missing observations in the *HMI* study. Each completed-data plot is from one randomly selected dataset.

Low Mutual Information (LMI) Study

We next consider a study where we treat EDUC and DIFFSENS as $X^{(F_0)}$, INCTOT and OCCSCORE as $Y^{(F_0)}$, PROPTX99, SEX, RACE, MARST, MIGRATE1, HCOVANY, and PWTYPE as $X^{(NF_0)}$, and the remaining variables as $Y^{(NF_0)}$. We again make 50% of values MCAR for variables in $\mathcal{F}_0$ and 1% of values MCAR for variables in $\mathcal{NF}_0$. The four variables in $\mathcal{F}_0$ frequently are used to assess socioeconomic status, which motivates why we create a simulation where they are the variables with high rates of missing data.

PROPTX99 has high relative mutual information with INCTOT and OCCSCORE as described previously. It also has relative mutual information values of 0.16 for EDUC and DIFFSENS, the two categorical focus variables. Other relationships are comparatively weak, with only one value exceeding 0.10 (AGE and DIFFSENS at 0.13). Thus, we add only PROPTX99 to the focus variables under *Move*.

Based on results in Section 3, we do not expect moving PROPTX99 to $\mathcal{F}$ to improve the quality of imputations substantially. In the simulations of Scenario 3 where we moved categorical variables highly associated with continuous $Y^{(F_0)}$, which most closely matches the characteristics of the *LMI* setting, *Move* and *Stay* had similar performances. The results from *LMI* bear this out. We compare the marginal probability densities of INCTOT and OCCSCORE, the marginal cell counts of EDUC and DIFFSENS, the joint distributions of (INCTOT, OCCSCORE), (INCTOT, PROPTX99), and (OCCSCORE, PROPTX99), and the associations of (INCTOT, EDUC), (OCCSCORE, EDUC), (PROPTX99, EDUC), (INCTOT, DIFFSENS), (OCCSCORE, DIFFSENS), and (PROPTX99, DIFFSENS). We find that *Stay* and *Move* perform very similarly. They also are not very different from HCMM-LD. To save space, we do not present these results here; details are in Chapter 4 of [20].

5. CONCLUSION

In general, the results of the artificial data simulations and the empirical study tell a consistent story. Compared to *Stay*, *Move* can improve estimation of the distribution of focus categorical variables, particularly for their associations with the variables moved to $\mathcal{F}$. *Move* improved the estimate of the association between INCTOT and PROPTX99, as well as OCCSCORE and PROPTX99, in *HMI*. The degree of improvement depends on the strength of the association between $X^{(F_0)}$ and $\mathcal{NF}_0$. This is evident in the result that *Move* did not substantially improve the accuracy of estimates involving EDUC in both *HMI* and *LMI*, as well as those involving DIFFSENS in *LMI*. For continuous variables in $\mathcal{F}_0$, *Stay* and *Move* performed similarly, suggesting that *Move* does not help much in terms of accuracy when the initial focus variables are continuous.

As a final comment, we note that *Move* and *Stay* can offer computational advantages over HCMM-LD. With HCMM-LD, one models all continuous variables with a multivariate normal distribution, which can result in a large number of covariance parameters when there are many continuous variables. In contrast, both *Stay* and *Move* assume that $Y^{(NF)}$ are locally independent, thereby removing them from the multivariate normal distributions.

ACKNOWLEDGMENTS

This work has been supported by the grant NSF SES 1131897.

REFERENCES

[1] Banerjee, A.; Murray, J. and Dunson, D.B. (2013). Bayesian learning of joint distributions of objects, *Journal of Machine Learning Research Workshop and Conference Proceedings*, **31**, 1–9.

[2] DeYoreo, M. and Kottas, A. (2015). A fully nonparametric modeling approach to binary regression, *Bayesian Analysis*, **10**, 821–847.

[3] DeYoreo, M. and Kottas, A. (2017). Bayesian nonparametric modeling for multivariate ordinal regression, *Journal of Computational and Graphical Statistics*, Forthcoming.

[4] DeYoreo, M.; Reiter, J.P. and Hillygus, D.S. (2017). Nonparametric Bayesian models with focused clustering for mixed ordinal and nominal data, *Bayesian Analysis*, **12**, 679–703.

[5] Dunson, D.B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data, *Journal of the American Statistical Association*, **104**, 1042–1051.

[6] Hannah, L.A.; Blei, D.M. and Powell, W.B. (2011). Dirichlet process mixtures of generalized linear models, *Journal of Machine Learning Research*, **12**, 1923–1953.

[7] Hu, J.; Reiter, J.P. and Wang, Q. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data, *Bayesian Analysis*, **13**, 183–200.

[8] Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, **96**, 161–173.

[9] Kim, H.J.; Reiter, J.P.; Wang, Q.; Cox, L.H. and Karr, A.F. (2014). Multiple imputation of missing or faulty values under linear constraints, *Journal of Business & Economic Statistics*, **32**, 375–386.

[10] Kim, H.J.; Cox, L.H.; Karr, A.F.; Reiter, J.P. and Wang, Q. (2015). Simultaneous edit-imputation for continuous microdata, *Journal of the American Statistical Association*, **110**, 987–999.

[11] Manrique-Vallier, D. and Reiter, J.P. (2013). Bayesian multiple imputation for large-scale categorical data with structural zeros, *Survey Methodology*, **40**, 125–134.

[12] Manrique-Vallier, D. and Reiter, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros, *Journal of Computational and Graphical Statistics*, **23**, 1061–1079.

[13] Müller, P. and Quintana, F.A. (2004). Nonparametric Bayesian data analysis, *Statistical Science*, **19**, 95–110.

[14] Müller, P. and Mitra, R. (2013). Bayesian nonparametric inference–why and how, *Bayesian Analysis*, **8**, 269–302.

[15] Murray, J.S. and Reiter, J.P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence, *Journal of the American Statistical Association*, **111**, 1466–1479.

[16] Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys, *Wiley Series in Probability and Mathematical Statistics: Applied probability and statistics*, Wiley & Sons, New York.

[17] Sethuraman, Jayaram (1994). A constructive definition of Dirichlet priors, *Statistica Sinica*, **4**, 639–650.

[18] Si, Y. and Reiter, J.P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys, *Journal of Educational and Behavioral Statistics*, **38**, 499–521.

[19] Wade, S.; Dunson, D.B.; Petrone, S. and Trippa, L. (2014). Improving prediction from Dirichlet process mixtures via enrichment, *Journal of Machine Learning Research*, **15**, 1041–1071.

[20] Wei, L. (2016). Methods for Imputing Missing Values and Synthesizing Confidential Values for Continuous and Magnitude Data, *Duke University*.

REVSTAT – Statistical Journal
Volume 16, Number 2, April 2018, 231–255

SEMI-PARAMETRIC LIKELIHOOD INFERENCE FOR BIRNBAUM–SAUNDERS FRAILTY MODEL

Authors: N. Balakrishnan
– Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
bala@mcmaster.ca

Kai Liu
– Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
liuk25@math.mcmaster.ca

Received: April 2017 Revised: August 2017 Accepted: August 2017

Abstract:

- Cluster failure time data are commonly encountered in survival analysis due to different factors such as shared environmental conditions and genetic similarity. In such cases, careful attention needs to be paid to the correlation among subjects within same clusters. In this paper, we study a frailty model based on Birnbaum–Saunders frailty distribution. We approximate the intractable integrals in the likelihood function by the use of Monte Carlo simulations and then use the piecewise constant baseline hazard function within the proportional hazards model in frailty framework. Thereafter, the maximum likelihood estimates are numerically determined. A simulation study is conducted to evaluate the performance of the proposed model and the method of inference. Finally, we apply this model to a real data set to analyze the effect of sublingual nitroglycerin and oral isosorbide dinitrate on angina pectoris of coronary heart disease patients and compare our results with those based on other frailty models considered earlier in the literature.

Key-Words:

- *Birnbaum–Saunders distribution; censored data; cluster time data; frailty model; Monte Carlo simulation; piecewise constant hazards.*

AMS Subject Classification:

- 62N02.

1. INTRODUCTION

It is of natural interest in medical or epidemiological studies to examine the effects of treatments. Proportional hazards model, proposed by Cox [5], is the most popular model for the analysis of such survival data which models the hazard function as

$$h(t) = h_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{x}),$$

where $t, \boldsymbol{x}$ and h_0 are the time to certain event, set of covariates and baseline hazard function, respectively. This model makes a critical assumption of independent observations from the subjects. However, correlation commonly exists in survival data due to shared environmental factors or genetic similarity. Therefore, neglecting this correlation may lead to biased results. A convenient choice for modeling these kinds of correlation in survival data is the frailty model. The terminology frailty was first introduced by Vaupel *et al.* [20], while accounting for the heterogeneity of individuals in distinct clusters. Generally speaking, the more frail an individual is, the earlier the event of interest will be. A shared frailty model introduces multiplicative random effects, which is referred to as the frailty term, in the proportional hazards model, and is defined as follows. Let $(t_{ij}, \delta_{ij}, \boldsymbol{x}_{ij}), i = 1, ..., n, j = 1, ..., m_i$, be the failure time, censoring indicator, and the covariate vector of the jth individual in the ith cluster, where δ_{ij} is 1 if t_{ij} is not censored and 0 otherwise. Let y_i be the frailty shared commonly by all the subjects in the ith cluster. Then, given y_i, t_{ij} are assumed to be independent with hazard function

$$h(t_{ij}|y_i) = y_i h_0(t_{ij}) \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij}). \tag{1.1}$$

The frailties y_i are assumed to be independent and identically distributed with a distribution, called the frailty distribution. The baseline hazard $h_0(t_{ij})$ is arbitrary. A common parametric choice of the baseline hazard is Weibull. Klein [13] proposed a non-parametric estimate of the cumulative hazard function of baseline distribution and then used a profile likelihood function.

The most prevailing choice of the frailty distribution is gamma distribution due to its mathematical simplicity and the mathematical tractability of ensuing inference [13]. It has a closed-form for the conditional likelihood function, given the observed data, so that EM algorithm can be applied effectively to obtain the maximum likelihood estimates. Another possibility is the positive stable distribution proposed by Hougaard [10]. Furthermore, Hougaard [11] derived power variance function from the positive stable distribution, which contains the preceding frailty distributions as special cases. All these distributions have simple Laplace transforms and therefore facilitates convenient computation of maximum likelihood estimates. However, there is no real biological reason for their use. Nevertheless, when the Laplace transform of the frailty distribution is unknown,

the likelihood function becomes intractable. Lognormal distribution is one such example. McGilchrist and Aisbett [16] developed a best linear unbiased prediction (BLUP) estimation method in this case of lognormal frailty model. Balakrishnan and Peng [2] proposed the generalized gamma frailty model since the generalized gamma distribution contains the gamma, Weibull, lognormal and exponential distributions all as special cases. Consequently, the generalized gamma frailty model becomes more flexible and tend to provide good fit to data as displayed by Balakrishnan and Peng [2].

The two-parameter Birnbaum–Saunders (BS) family of distributions was originally derived as a fatigue model by Birnbaum and Saunders [3] for which a more general derivation from a biological viewpoint was later provided by Desmond [7]. This distribution possesses many interesting distributional properties and shape characteristics. In the present work, we use this BS model as the frailty distribution along with a piecewise constant baseline hazard function within the proportional hazards model to come up with a flexible frailty model. The precise specification of this model is detailed in Section 2. An estimation method to obtain the maximum likelihood estimates of model parameters is presented in Section 3. A simulation study is conducted in Section 4 to assess the performance of the proposed method and then the usefulness of the proposed model and the method of inference is illustrated with a real data in Section 5. Discussions and some concluding remarks are finally made in Section 6.

2. MODEL SPECIFICATION

2.1. BS distribution as frailty distribution

The BS distribution was originally derived to model fatigue failure caused under cyclic loading [3]. The fatigue failure is due to the initiation, growth and ultimate extension of a dominant crack. It is assumed that the total crack extension Y_j due to the jth cycle, for $j = 1, ...,$ are independent and identically distributed random variables with mean μ and variance σ^2. Then, the distribution of the failure time (i.e., time for the crack to exceed a certain threshold level) is given by

$$F(t;\alpha,\beta) = \Phi\left[\frac{1}{\alpha}\left\{\left(\frac{t}{\beta}\right)^{1/2} - \left(\frac{\beta}{t}\right)^{1/2}\right\}\right], \qquad 0 < t < \infty, \quad \alpha, \beta > 0, \tag{2.1}$$

where Φ is the standard normal cumulative distribution function (CDF), and α and β are the shape and scale parameters, respectively. We now assume that the frailty random variable Y_i in (1.1) follows the BS distribution defined in (2.1).

Since $\frac{1}{\alpha}\left\{\left(\frac{T}{\beta}\right)^{1/2}-\left(\frac{\beta}{T}\right)^{1/2}\right\}$ is a standard normal random variable, the random variable T is simply given by

$$T = \beta\left\{\frac{\alpha Z}{2}+\left[\left(\frac{\alpha Z}{2}\right)^2+1\right]^{1/2}\right\}^2, \tag{2.2}$$

where $Z \sim N(0,1)$. The probability density function (PDF) of T, derived from (2.1), is given by

$$f(t;\alpha,\beta) = \frac{1}{2\sqrt{2\pi}\,\alpha\beta}\left[\left(\frac{\beta}{t}\right)^{1/2}+\left(\frac{\beta}{t}\right)^{3/2}\right]\exp\left[-\frac{1}{2\alpha^2}\left(\frac{t}{\beta}+\frac{\beta}{t}-2\right)\right], \quad t>0. \tag{2.3}$$

The relation between T and Z in (2.2) enables us to obtain the mean and variance of T easily as

$$E(T) = \beta\left(1+\frac{1}{2}\,\alpha^2\right), \tag{2.4}$$

$$V(T) = (\alpha\beta)^2\left(1+\frac{5}{4}\,\alpha^2\right). \tag{2.5}$$

In the frailty model in (1.1), if the frailty term y_i is assumed to follow the BS distribution, for ensuring identifiability of model parameters, the mean of the frailty distribution needs to be set as 1. More specifically, let Y_1 be a BS random variable with shape parameter α and scale parameter β with its mean as 1. Let $Y_2 = cY_1$. Then, $E(Y_2) = cE(Y_1) = c$. Besides, we know that if $Y_1 \sim \mathrm{BS}(\alpha,\beta)$, then $cY_1 \sim \mathrm{BS}(\alpha, c\beta)$. Therefore, $Y_2 \sim \mathrm{BS}(\alpha, c\beta)$ with mean c. Then, given the frailty term y_2, the lifetime of the patients are modeled by the hazard function

$$h(t|y_2) \;=\; y_2\, h_0(t)\exp(\boldsymbol{\beta'x}) \;=\; c\, y_1\, h_0(t)\exp(\boldsymbol{\beta'x}).$$

Let us define $ch_0(t)$ to be a new baseline hazard function $h_1(t)$, which is nothing but rescaling the original baseline hazard function. Then, the model can be rewritten as

$$h(t|y_2) \;=\; y_1\, h_1(t)\exp(\boldsymbol{\beta'x}),$$

which is identical to a frailty model with frailty variable Y_1 and baseline hazard function $h_1(t) = ch_0(t)$.

Thus, the scale parameter β can be written in terms of the shape parameter α as

$$\beta = \frac{2}{2+\alpha^2}, \tag{2.6}$$

so that the variance of the frailty variable Y_i becomes

$$V(Y_i) = \frac{4\alpha^2+5\alpha^4}{\alpha^4+4\alpha^2+4}, \tag{2.7}$$

which is constrained to be in the interval $(0,5)$.

Some important discussions on inferential issues for BS distribution can be found in [1, 4, 8, 9, 15, 17, 18, 19].

2.2. Piecewise constant hazard as baseline hazard function

The baseline hazard $h_0(t)$ in (1.1) is normally assumed in the parametric setting to be that of exponential or Weibull distribution [11]. However, such a strong parametric assumption is not always desirable as the resulting inference may become non-robust. For this reason, we use a piecewise constant hazard function to approximate the baseline hazard so that it could capture inherent shape and features of the hazard function better. Let J be the number of partitions of the time interval, i.e., $0 = t^{(0)} < t^{(1)} < \cdots < t^{(J)}$, where $t^{(J)} > \max(t_{ij})$. The points $t^{(1)}, ..., t^{(J)}$ are called cut-points. The piecewise constant hazard function is then given by

$$h_0(t) = \gamma_k \qquad \text{for} \quad t^{(k-1)} \le t < t^{(k)} \quad \text{for} \quad k = 1, ..., J.$$

The corresponding cumulative hazard function is

$$H_0(t) = \sum_{q=1}^{k-1} \gamma_q \Big(t^{(q)} - t^{(q-1)} \Big) + \gamma_k \Big(t - t^{(k-1)} \Big) \qquad \text{for} \quad t^{(k-1)} \le t < t^{(k)}, \tag{2.8}$$

where γ_k is a constant hazard for interval $\big[t^{(k-1)}, t^{(k)}\big)$, $k = 1, ..., J$.

3. ESTIMATION METHOD

Let $(t_{ij}, \delta_{ij}, \boldsymbol{x}_{ij}), i = 1, ..., n, j = 1, ..., m_i$, be the failure time, censoring indicator, and the covariate vector for the jth individual in the ith cluster and y_i be the frailty term. Then, the full likelihood function of the BS frailty model is obtained from (1.1) as

$$\begin{aligned}
L &= \prod_{i=1}^{n} \int_0^\infty \Big(\prod_{j=1}^{m_i} h(t_{ij}|y_i)^{\delta_{ij}} S(t_{ij}|y_i) \Big) f(y_i)\, dy_i \\
&= \prod_{i=1}^{n} \int_0^\infty \Big[\prod_{j=1}^{m_i} \Big(y_i h_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}) \Big)^{\delta_{ij}} \\
&\quad \times \exp\Big(- y_i H_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}) \Big) \Big] f(y_i)\, dy_i \\
&= \prod_{i=1}^{n} \Big[\prod_{j=1}^{m_i} \Big(h_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}) \Big)^{\delta_{ij}} \\
&\quad \times \int_0^\infty y_i^{\delta_{i\cdot}} \exp\Big(- y_i \sum_{j=1}^{m_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}) \Big) f(y_i)\, dy_i \Big] \\
&= \prod_{i=1}^{n} \Big[\prod_{j=1}^{m_i} \Big(h_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}) \Big)^{\delta_{ij}} I_i \Big],
\end{aligned} \tag{3.1}$$

where $\delta_{i\cdot} = \sum_{j=1}^{m_i} \delta_{ij}$, H_0 is the cumulative baseline hazard function with parameter $\boldsymbol{\gamma}$ as given in (2.8), f is the PDF of the BS distribution with shape parameter α and scale parameter $\beta = \frac{2}{2+\alpha^2}$ as given in (2.3), and

$$I_i = \int_0^\infty y_i^{\delta_{i\cdot}} \exp\Big(- y_i \sum_{j=1}^{m_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big) f(y_i)\, dy_i.$$

The above expression of I_i can be rewritten as

$$I_i = \int_{-\infty}^\infty g(z_i)^{\delta_{i\cdot}} \exp\Big(- g(z_i) \sum_{j=1}^{m_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big) f_Z(z_i)\, dz_i, \tag{3.2}$$

where f_Z is the PDF of the standard normal distribution and

$$g(z_i) = \frac{2}{2+\alpha^2}\left\{1 + \frac{\alpha^2 z_i^2}{2} + \alpha\, z_i\Big(1 + \frac{\alpha^2 z_i^2}{4}\Big)^{1/2}\right\}.$$

The maximum likelihood estimates are hard to determine due to the intractable integral in (3.2) present in the likelihood function in (3.1). A direct and convenient way is to use Monte Carlo simulation to approximate the integral in (3.2) as follows:

$$\begin{aligned} I_i &= E_Z\Big[g(Z)^{\delta_{i\cdot}} \exp\Big(- g(Z) \sum_{j=1}^{m_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big)\Big] \\ &= \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}} \exp\Big(- g(z_{(k)}) \sum_{j=1}^{m_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big), \end{aligned}$$

where $z_{(k)}, k = 1, ..., N$, are the realizations of standard normal random variable.

The log-likelihood function can then be approximated from (3.1) as

$$\begin{aligned} l = \sum_{i=1}^{n} \Bigg[& \sum_{j=1}^{m_i} \delta_{ij}\Big(\log h_0(t_{ij}) + \boldsymbol{\beta}'\boldsymbol{x}_{ij}\Big) \\ & + \log \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}} \exp\Big(- g(z_{(k)}) \sum_{j=1}^{m_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big)\Bigg]. \end{aligned} \tag{3.3}$$

Once the approximate log-likelihood function is obtained as in (3.3), Fisher's score function and the Hessian matrix with respect to the parameters $\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}$ can be obtained readily upon taking partial derivatives of first- and second-order, and pertinent details are presented in Appendix A. The MLEs of model parameters can then be obtained by Newton–Raphson algorithm iteratively as

$$\begin{bmatrix} \hat\alpha^{(k)} \\ \hat{\boldsymbol{\beta}}^{(k)} \\ \hat{\boldsymbol{\gamma}}^{(k)} \end{bmatrix} = \begin{bmatrix} \hat\alpha^{(k-1)} \\ \hat{\boldsymbol{\beta}}^{(k-1)} \\ \hat{\boldsymbol{\gamma}}^{(k-1)} \end{bmatrix} - \begin{bmatrix} \frac{\partial^2 l}{\partial\alpha^2} & \frac{\partial^2 l}{\partial\alpha\partial\boldsymbol{\beta}^T} & \frac{\partial^2 l}{\partial\alpha\partial\boldsymbol{\gamma}^T} \\ \frac{\partial^2 l}{\partial\alpha\partial\boldsymbol{\beta}} & \frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^T} \\ \frac{\partial^2 l}{\partial\alpha\partial\boldsymbol{\gamma}} & \frac{\partial^2 l}{\partial\boldsymbol{\beta}^T\partial\boldsymbol{\gamma}} & \frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l}{\partial\alpha} \\ \frac{\partial l}{\partial\boldsymbol{\beta}} \\ \frac{\partial l}{\partial\boldsymbol{\gamma}} \end{bmatrix}_{\alpha=\hat\alpha^{(k-1)}, \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k-1)}, \boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}^{(k-1)}}.$$

The iterations need to be continued until the desired tolerance level is achieved, say, $|\hat{\boldsymbol{\theta}}_{i+1} - \hat{\boldsymbol{\theta}}_i| < 10^{-6}$. Finally, the standard errors of the estimates of $\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}$ can be obtained from the inverse of the Hessian matrix evaluated at the determined MLEs.

4. SIMULATION STUDY

An extensive simulation study is carried out here to assess the performance of the proposed model and the method of estimation. We consider 4 scenarios: (1) $n = 100,\ m = 2$, (2) $n = 100,\ m = 4$, (3) $n = 100,\ m = 8$ and (4) $n = 400,\ m = 2$. Here, the clusters can be considered as hospitals and each subject as a patient in these hospitals. The patients are randomly assigned to either a treatment group or a control group with equal probability. The frailty term follows (1) the BS distribution with shape parameter $\frac{(2\sqrt{10}-2)^{1/2}}{3}$ and scale parameter $\frac{9}{8+\sqrt{10}}$, (2) gamma distribution (GA) with shape parameter 2 and scale parameter 0.5, (3) lognormal (LN) distribution with $\mu = -\frac{\log(1.5)}{2}$ and $\sigma^2 = \log(1.5)$. With these choices of parameters, the mean and variance of the frailty distribution become 1 and 0.5, respectively, for all these frailty distributions. The standard exponential distribution and the standard lognormal distribution are considered for baseline distributions. We then set $\beta = -\log(2) = -0.6931$ so that the hazard rate of patients in the treatment group is half of those in the control group. Finally, the censoring times are generated from the uniform distribution in $[0, 4.5]$.

The simulation procedure is as follows:

(**1**) Generate n frailty values from frailty distributions, i.e., $y_i, i = 1, ..., n$, and assign each subject in the same cluster with same frailty value.

(**2**) Assign each patient to treatment group or control group with probability 0.5.

(**3**) Given the frailty term, the survival function is

$$S(t_{ij}|y_i) = \exp\big(-y_i H_0(t_{ij}) \exp(\beta x_{ij})\big)$$

and the cumulative distribution function is

$$F(t_{ij}|y_i) = 1 - \exp\big(-y_i H_0(t_{ij}) \exp(\beta x_{ij})\big),$$

which follows a uniform distribution (0,1). Therefore we generate u_{ij} from Uniform(0,1) and set $F(t_{ij}|y_i) = u_{ij}$.

(**4**) Calculate the baseline cumulative hazard function, which is

$$H_0(t_{ij}) = -\frac{\log(1 - u_{ij})}{y_i \exp(\beta x_{ij})}.$$

(**5**) Solve for the lifetime according to the true baseline distribution, i.e.: for standard exponential, $t_{ij} = H_0(t_{ij})$; for standard lognormal, $t_{ij} = \exp\big(\Phi^{-1}(1 - \exp(-H_0(t_{ij})))\big)$ since $1 - \exp(-H_0(t_{ij})) = \Phi(\log(t_{ij}))$.

(**6**) Now, we generate censoring time c_{ij} from Uniform[0,4.5].

(**7**) Compare t_{ij} and c_{ij}. If $t_{ij} <= c_{ij}$, then set t_{ij} to be the observed time and the censoring indicator $\delta_{ij} = 1$. If $t_{ij} > c_{ij}$, we set c_{ij} to be our observed time and $\delta_{ij} = 0$.

We generated 1000 data sets under each setting and applied the proposed semi-parametric BS frailty model to these data sets. For comparative purposes, we fitted the simulated data sets with the parametric BS frailty model along with gamma and lognormal frailty models. Thus, we fitted 6 models for each simulated data with frailty distribution to be one of BS, gamma or lognormal, and the baseline hazard function to be either piecewise constant hazard function or Weibull hazard function. The primary parameters of interest are the treatment effect and the frailty variance, and so our attention will focus on these parameters. The estimates of the treatment effect are summarized in Figures 1 and 2, while Figures 3 and 4 demonstrate how the estimates of the frailty variance differ under different models. The horizontal black lines are the true values of the parameters of interest, while the vertical bars give 95% confidence intervals. The three numbers on the top of each plot are the rejection rate and coverage probabilities at confidence levels of 95% and 90%. The two numbers at the bottom of each plot provide bias and mean square error for the different models considered.

Figures 1 and 2 clearly show that the choice of frailty distribution has little impact on the estimate of treatment effect. When the true baseline distribution is exponential, either Weibull baseline hazard or piecewise constant hazard function will result in accurate estimation of the treatment effect. However, when the true baseline distribution is lognormal, use of piecewise constant hazard baseline distribution results in smaller bias and mean square error than when using the Weibull distribution as baseline. This reveals that misspecification of the baseline hazard function impacts the estimate of treatment effect and the semi-parametric frailty models are therefore better than the parametric frailty models based on robustness consideration.

The heterogeneity among clusters is explained by the frailty variance and so it is important to investigate the frailty variance. The estimates of frailty variance are shown in Figures 3 and 4. BS frailty model always has less mean square error than the lognormal frailty model no matter what the true frailty model is. Even though the gamma frailty model generally has smallest bias and mean square error, its coverage probabilities are quite small and considerably below the nominal level. Both parametric and semi-parametric BS frailty models have coverage probabilities close to the nominal level, and so does the lognormal frailty model. Furthermore, as the sample size gets larger, the estimates become more precise. When the sample size is small, the rejection rate is small for BS and lognormal frailty models, but they become larger when the sample size increases.

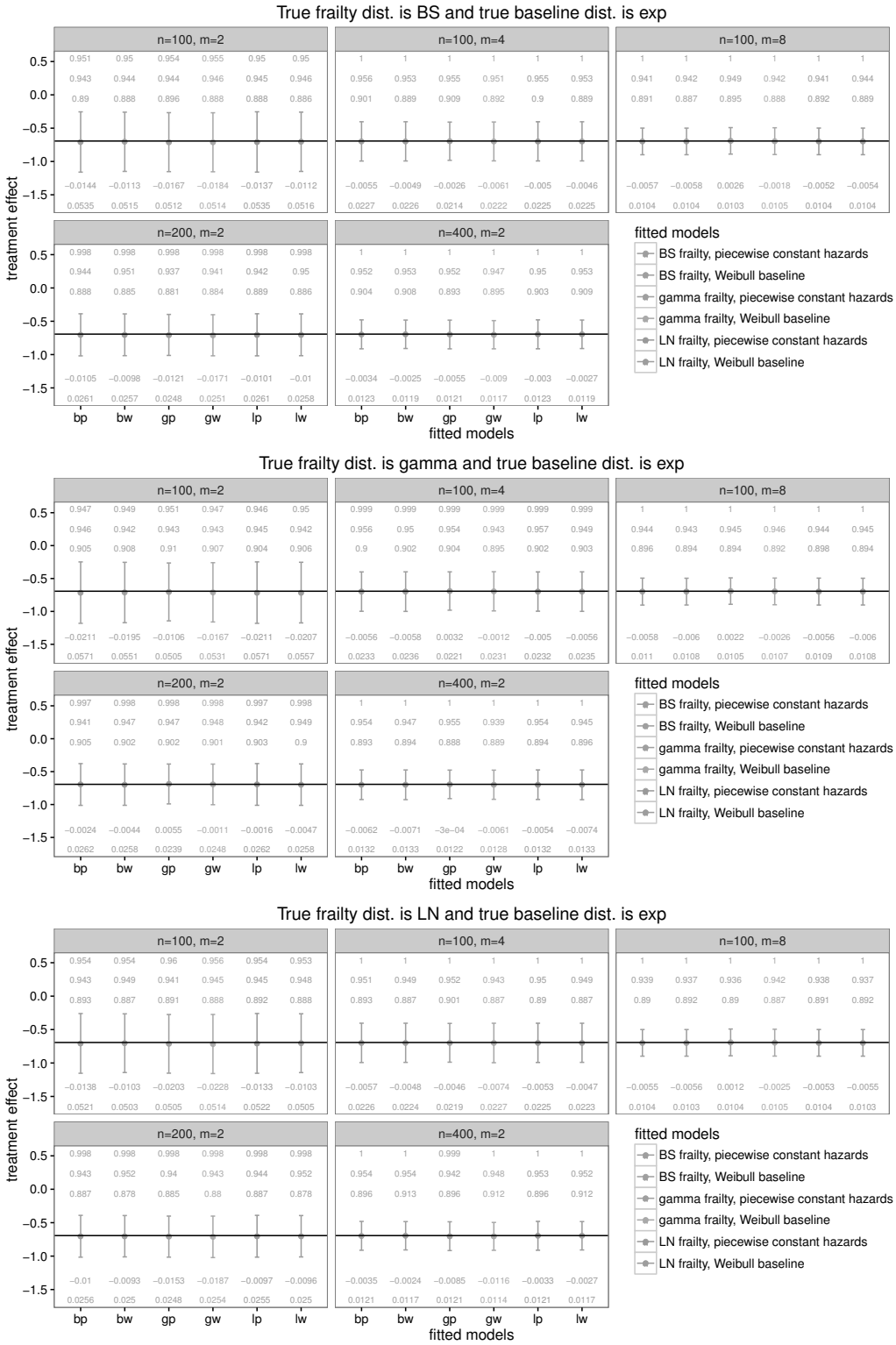


Figure 1: Estimate of treatment effect when the true baseline distribution is exponential.

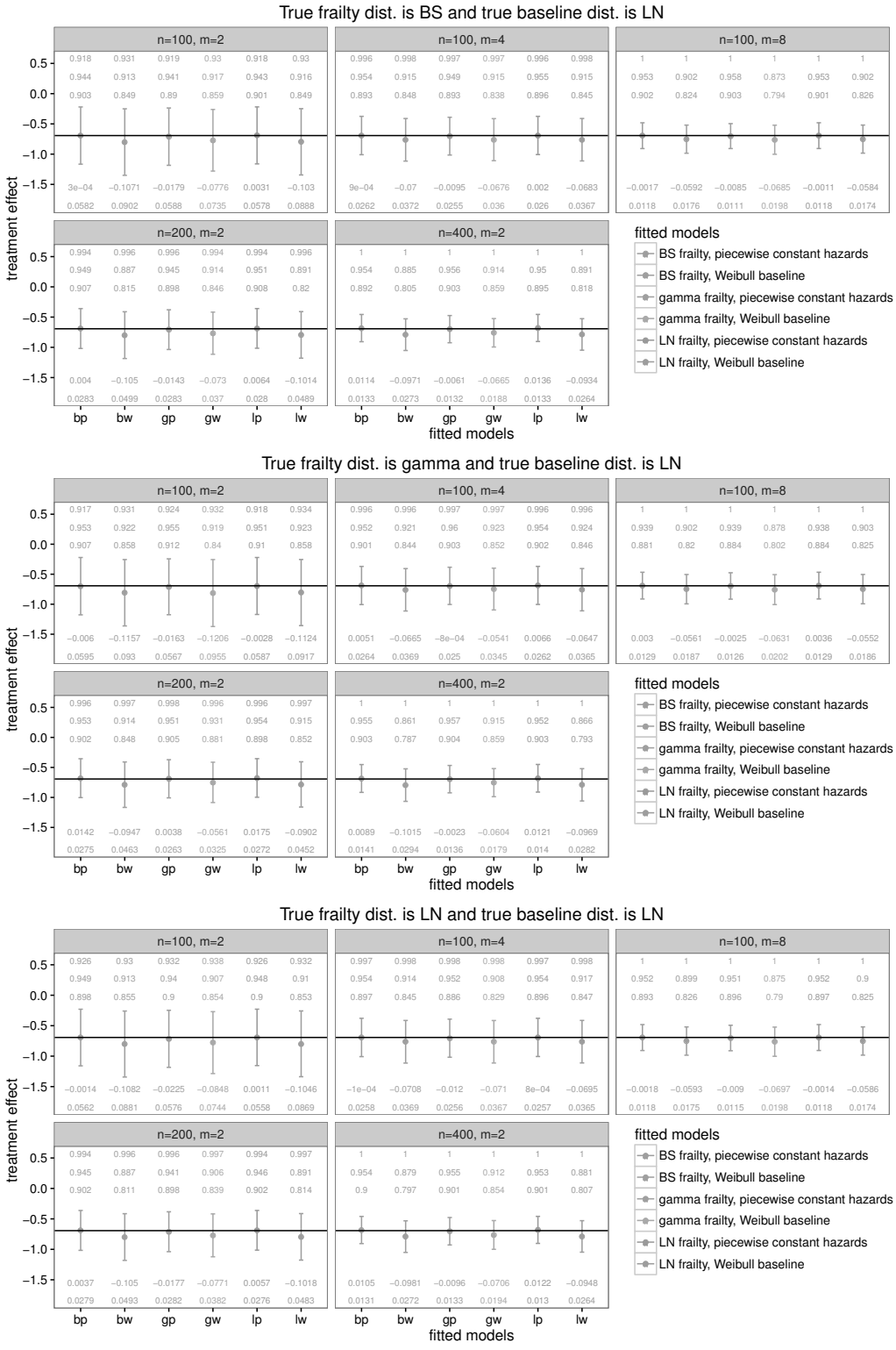


Figure 2: Estimate of treatment effect when the true baseline distribution is lognormal.

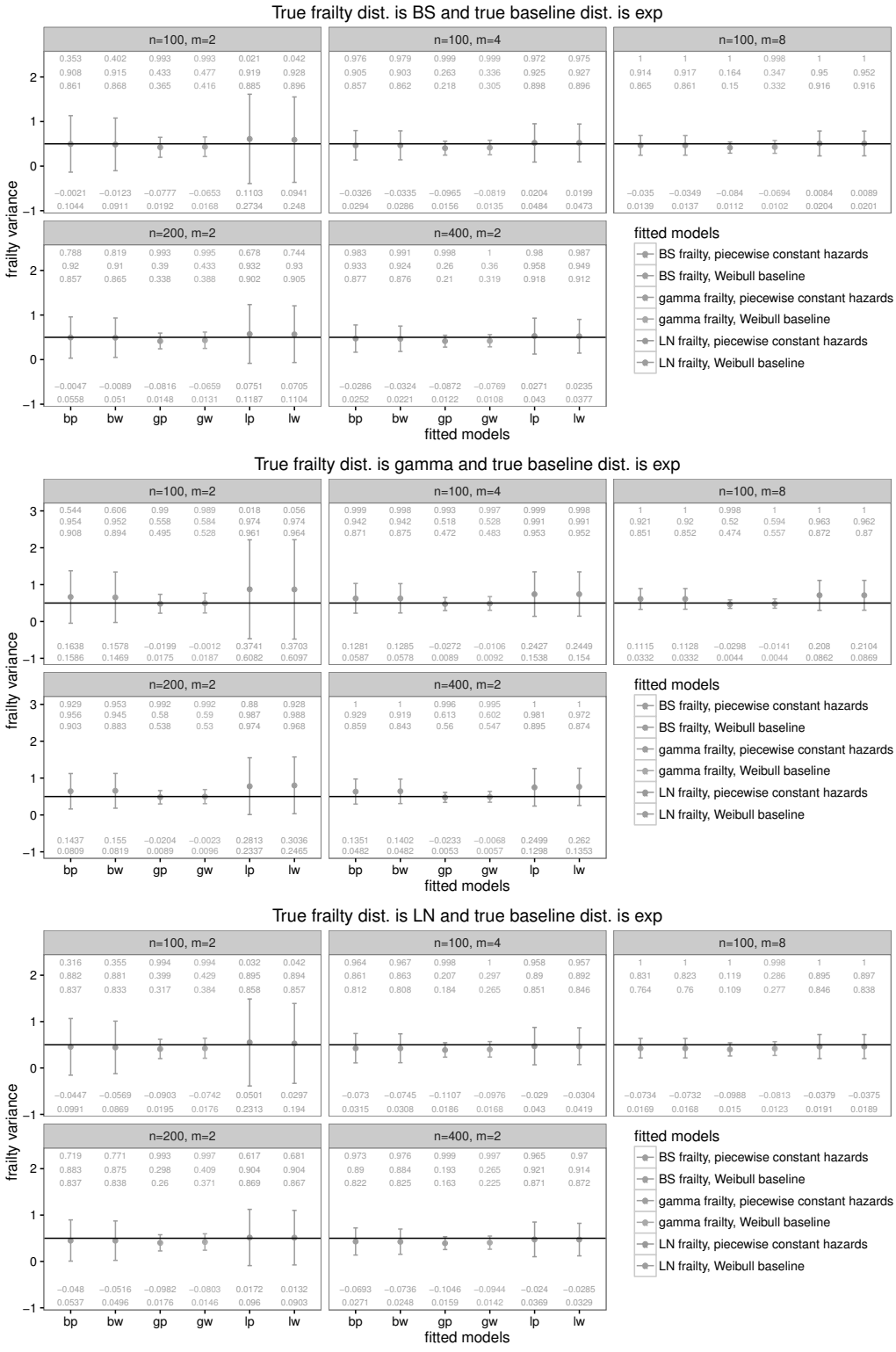


Figure 3: Estimate of frailty variance when the true baseline distribution is exponential.

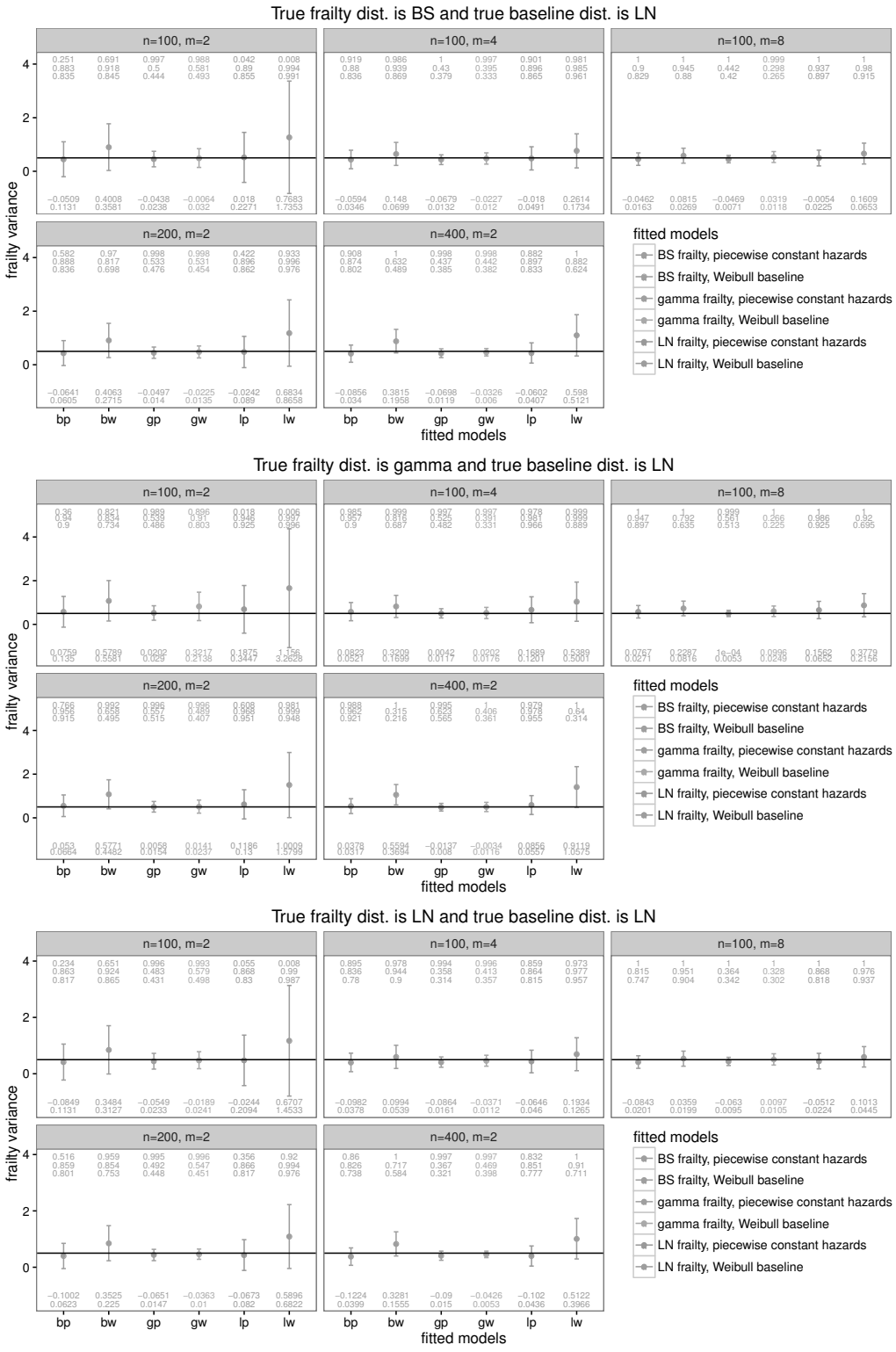


Figure 4: Estimate of frailty variance when the true baseline distribution is lognormal.

Table 1 summarizes the selection rate of the models based on the log-likelihood value. When the true baseline distribution is exponential, the models with correct frailty distributions generally have the largest selection rate except when the true frailty distribution is lognormal and the number of clusters is 100.

Table 1: Observed selection rates based on log-likelihood value.

Fitted models		True models					
baseline	frailty	BS		GA		LN	
		Exp	LN	Exp	LN	Exp	LN
$n = 100,\ m = 2$							
Weibull	BS	0.385	0.587	0.285	0.296	0.399	0.580
	GA	0.262	0.283	0.418	0.633	0.232	0.280
	LN	0.353	0.130	0.297	0.071	0.369	0.140
Piecewise	BS	0.378	0.500	0.282	0.385	0.394	0.532
	GA	0.242	0.325	0.400	0.489	0.224	0.294
	LN	0.380	0.175	0.318	0.126	0.382	0.174
$n = 100, m = 4$							
Weibull	BS	0.386	0.512	0.227	0.435	0.392	0.506
	GA	0.287	0.322	0.563	0.463	0.249	0.305
	LN	0.327	0.166	0.210	0.102	0.359	0.189
Piecewise	BS	0.385	0.449	0.224	0.289	0.381	0.455
	GA	0.282	0.338	0.571	0.590	0.244	0.286
	LN	0.333	0.213	0.205	0.121	0.375	0.259
$n = 100, m = 8$							
Weibull	BS	0.518	0.594	0.189	0.318	0.467	0.539
	GA	0.203	0.190	0.676	0.579	0.162	0.194
	LN	0.279	0.216	0.135	0.103	0.371	0.267
Piecewise	BS	0.510	0.576	0.204	0.209	0.439	0.497
	GA	0.203	0.225	0.666	0.690	0.186	0.202
	LN	0.287	0.199	0.130	0.101	0.375	0.301
$n = 200, m = 2$							
Weibull	BS	0.362	0.668	0.251	0.673	0.337	0.642
	GA	0.254	0.275	0.508	0.293	0.251	0.290
	LN	0.284	0.057	0.241	0.034	0.412	0.068
Piecewise	BS	0.339	0.518	0.268	0.356	0.354	0.548
	GA	0.284	0.376	0.501	0.597	0.250	0.334
	LN	0.377	0.106	0.231	0.047	0.396	0.118
$n = 400, m = 2$							
Weibull	BS	0.365	0.755	0.221	0.774	0.345	0.714
	GA	0.237	0.237	0.561	0.217	0.214	0.272
	LN	0.398	0.008	0.218	0.009	0.441	0.014
Piecewise	BS	0.340	0.545	0.221	0.285	0.339	0.594
	GA	0.239	0.413	0.546	0.697	0.228	0.352
	LN	0.421	0.042	0.233	0.018	0.433	0.054

Under this situation, the BS frailty models have slightly greater selection rates than the lognormal frailty model. In fact, the log-likelihood values are quite close for BS and lognormal frailty models. When the number of clusters increases, the selection rate of the lognormal frailty model increases and indeed becomes the largest in the case when the true frailty distribution is lognormal. On the other hand, when the true baseline distribution is lognormal, the parametric BS frailty model becomes more likely to be selected, especially when the number of clusters increases. Use of the piecewise constant hazard baseline function results in increasing selection probability of the true frailty distribution when the frailty distribution is gamma. However, the semi-parametric BS frailty model often has the highest selection rate when the true frailty distribution is lognormal. This suggests that the semi-parametric BS frailty model often results in MLEs with larger likelihood values than the semi-parametric lognormal frailty model and thus provide a better fit to observed data.

In summary, the choice of frailty distribution and the baseline distribution is a critical issue in frailty modeling. An inappropriate baseline distribution seems lead to larger errors in the estimation of both treatment effect and the frailty variance. However, the choice of the frailty distribution has less influence on estimating the treatment effect, but it highly impacts the estimation of frailty variance. Finally, the proposed BS frailty model provides a robust estimate of treatment effect and the frailty variance overall, and generally results in larger likelihood values among all fitted models. The R codes are available upon request from the authors.

5. ILLUSTRATION WITH A CORONARY HEART DISEASE STUDY

In this section, we fit the proposed semi-parametric BS frailty model to a real data set from Danahy *et al.* [6] concerning a study of oral administration of isosorbide dinitrate on 21 coronary heart disease patients, presented in Table 2. In the study, the patients were treated initially with sublingual nitroglycerin (SLN) and sublingual placebo (SLP) and then two tests of bike pedalling were conducted on the patients. Then, they took oral isosorbide dinitrate (OI) and oral placebo (OP) after which eight bike pedalling tests were given right after (OI0, OP0) and 1h (OI1, OP1), 3h (OI3, OP3), 5h (OI5, OP5). The times to angina pectoris were then recorded. Some of the times were censored because the patients were too exhausted (times with * are the censoring times).

Hougaard [12] studied the effects of the treatments with the proportional hazards model. In addition, several frailty models with gamma, stable and power variance function as the frailty distribution, along with non-parametric and Weibull hazard functions, were fitted to these data. The analyses carried

out demonstrated that a frailty model fitted the data better than the classical proportional hazards model, and the power variance function frailty distribution was more suitable than the gamma frailty distribution. Balakrishnan and Peng [2] analyzed the same data with a generalized gamma frailty model (GG) with both parametric and semi-parametric baseline hazard functions. These authors showed that the generalized gamma frailty model provided a better fit than the gamma, Weibull and lognormal frailty models, which are all special cases of the generalized gamma frailty model.

Table 2: Exercise times to Angina Pectoris (in seconds).

ID	SLP	SLN	OP0	OP1	OP3	OP5	OI0	OI1	OI3	OI5
1	155	431	150	172	118	143	136	445*	393*	226
2	269	259	205	287	211	207	250	306	206	224
3	408	446	221	244	147	250	215	232	258	268
4	308	349	150	290	205	210	235	248	298	207
5	135	175	87	157	135	105	129	121	110	102
6	409	523	301	357	388	388	425	580	613	514
7	455	488	342	390	441	468	441	504*	519*	484*
8	182	227	215	210	188	189	208	264	210	172
9	141	102	131	125	99	115	154	110	123	105
10	104	231	108	114	136	111	89	145	172	123
11	207	249	228	224	251	206	250	230	264	216
12	198	247	190	199	243	222	147	403	290	208
13	274	397	234	249	267	241	231	540*	370	316
14	191	251	218	194	197	223	224	432	291	212
15	156	401	199	329	197	176	152	733*	492	303
16	458	766	406	431	448	328	417	743*	566	391
17	188	199	194	168	168	159	213	250	150	180
18	258	566*	277	264	276	251	490	559*	557*	439
19	437	552	424	512	560	478	406	651	624	554
20	115	237	234	232	281	237	229	327	280	321
21	200	387	227	199	223	227	265	565*	504*	517*

We first investigate the feature of the data through the cumulative hazard plot, presented in Figure 5. The cumulative hazard after taking placebo is seen to be higher than that after taking sublingual nitroglycerin or isosorbide dinitrate. The hazard rate is increasing after taking placebo while it looks to be increasing and then decreasing after taking isosorbide dinitrate. We then fitted these data with the parametric and semi-parametric BS frailty models. The obtained results are presented in Tables 3 and 4. In addition, we fitted the parametric and semi-parametric gamma (GA), lognormal (LN) and inverse Gaussian (IG) frailty models to these data. Furthermore, for comparative purpose, we also include estimates of the generalized gamma frailty model (GG) from [2]. The minimum and maximum time to angina pectoris were 87s and 766s, respectively. Figure 6 is a histogram of observed times and it

shows that the data is sparse at the tail. So, we chose the cutpoints to be $t^{(0)} = 87, t^{(1)} = 150, t^{(2)} = 200, t^{(3)} = 250, t^{(4)} = 300, t^{(5)} = 400, t^{(6)} = 766$ to capture changes in the piecewise constant hazard baseline function.

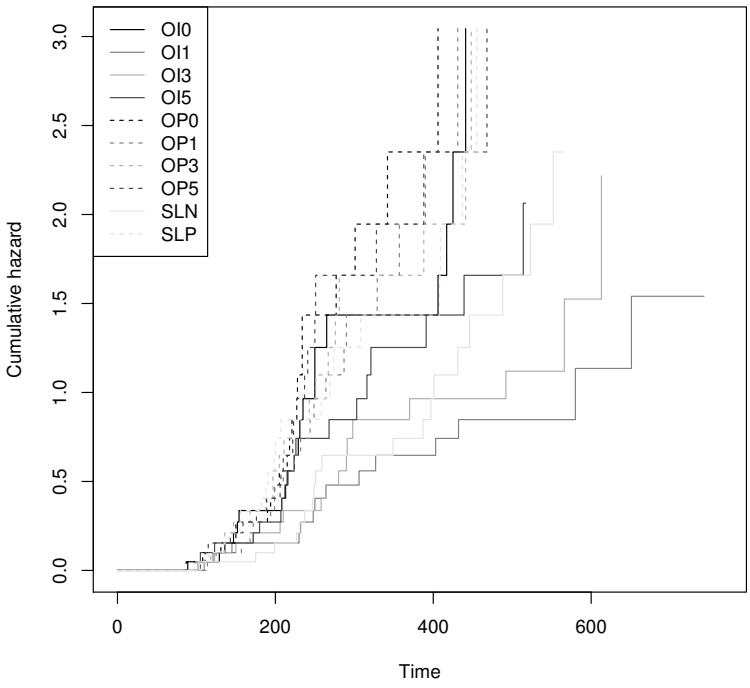


Figure 5: Cumulative hazard plot of the treatments.

Table 3: Fitted frailty models with Weibull baseline hazard function.

	BS	GA	LN	IG	GG
SLN	−1.54(0.34)	−1.51(0.34)	−1.55(0.34)	−1.43(0.34)	−1.51(0.34)
OP0	0.69(0.33)	0.69(0.33)	0.69(0.33)	0.7(0.33)	0.67(0.33)
OP1	0.12(0.32)	0.13(0.32)	0.12(0.32)	0.17(0.32)	0.11(0.33)
OP3	0.26(0.33)	0.28(0.33)	0.26(0.33)	0.24(0.33)	0.27(0.32)
OP5	0.63(0.33)	0.64(0.33)	0.63(0.33)	0.57(0.33)	0.66(0.32)
OI0	0.13(0.33)	0.15(0.33)	0.13(0.33)	0.15(0.33)	0.19(0.32)
OI1	−2.67(0.41)	−2.64(0.41)	−2.68(0.41)	−2.55(0.4)	−2.54(0.39)
OI3	−1.38(0.36)	−1.37(0.36)	−1.40(0.36)	−1.36(0.36)	−1.31(0.35)
OI5	−0.35(0.35)	−0.35(0.35)	−0.37(0.35)	−0.41(0.35)	−0.38(0.33)
$\log(p)$	1.59(0.06)	1.58(0.05)	1.59(0.06)	1.56(0.05)	1.59(0.06)
$\log(\lambda)$	−25.66(1.53)	−25.44(1.44)	−25.49(1.66)	−25.34(1.58)	−25.40(8.14)
Frailty variance	3.35(0.45)	2.51(0.07)	49.33(54.78)	10.87(1.19)	232.27(617.98)
Log−likelihood	−1121.86	−1124.86	−1122.12	−1123.00	1120.92
AIC	2267.71	2273.72	2268.23	2270.00	2267.82

Table 4: Fitted frailty models with piecewise constant baseline hazard function.

	BS	GA	LN	IG	GG
SLN	−1.38(0.34)	−1.38(0.33)	−1.38(0.34)	−1.43(0.33)	−1.37(0.34)
OP0	0.57(0.33)	0.59(0.33)	0.58(0.33)	0.50(0.32)	0.56(0.33)
OP1	−0.004(0.32)	0.01(0.32)	−0.01(0.32)	−0.07(0.31)	−0.06(0.31)
OP3	0.20(0.33)	0.22(0.32)	0.20(0.33)	0.12(0.32)	0.17(0.33)
OP5	0.50(0.32)	0.52(0.32)	0.50(0.32)	0.42(0.32)	0.48(0.32)
OI0	0.05(0.32)	0.06(0.32)	0.05(0.32)	−0.02(0.32)	0.09(0.32)
OI1	−2.18(0.38)	−2.24(0.38)	−2.17(0.39)	−2.21(0.38)	−2.16(0.38)
OI3	−1.29(0.35)	−1.32(0.35)	−1.29(0.35)	−1.34(0.35)	−1.28(0.35)
OI5	−0.39(0.34)	−0.41(0.34)	−0.40(0.34)	−0.46(0.33)	−0.41(0.41)
$\log(\gamma_1)$	−5.42(0.42)	−5.37(0.43)	−5.64(0.77)	−5.48(0.48)	−4.76(0.49)
$\log(\gamma_2)$	−3.96(0.48)	−3.95(0.46)	−4.15(0.76)	−3.98(0.44)	−3.24(0.48)
$\log(\gamma_3)$	−2.49(0.49)	−2.48(0.46)	−2.70(0.78)	−2.52(0.42)	−1.79(0.51)
$\log(\gamma_4)$	−1.83(0.52)	−1.79(0.49)	−2.05(0.80)	−1.91(0.42)	−1.14(0.57)
$\log(\gamma_5)$	−2.08(0.55)	−1.98(0.50)	−2.30(0.83)	−2.19(0.43)	−1.38(0.61)
$\log(\gamma_6)$	−0.57(0.56)	−0.43(0.50)	−0.79(0.84)	−0.71(0.41)	0.08(0.62)
Frailty variance	3.02(0.47)	2.34(0.07)	16.67(17.45)	10.52(1.17)	56.18(105.16)
Log−likelihood	−1111.88	−1116.58	−1112.16	−1111.562	−1111.39
AIC	2255.75	2265.16	2256.31	2255.12	2256.78

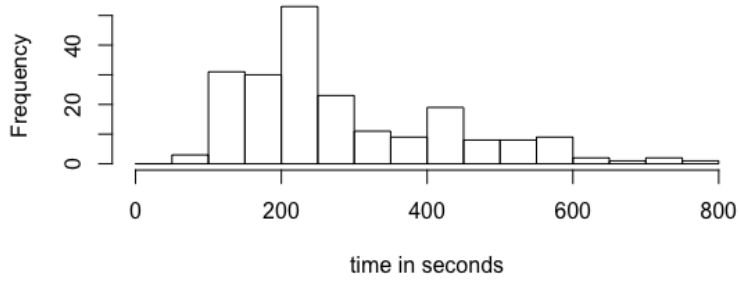


Figure 6: Histogram of observed times.

All the models result in similar estimates of the treatment effects, which are consistent with the results of Hougaard [12] and Balakrishnan and Peng [2]. Among all the frailty models fitted, the parametric BS frailty models provided the best fit since they had the smallest AIC values compared to other parametric models, even compared to the parametric generalized gamma frailty model possessing one extra shape parameter. Among the semi-parametric models, even though the inverse Gaussian model has the smallest AIC, the AIC of semi-parametric BS frailty model is quite close. Upon comparing the parametric and semi-parametric frailty models, we note that the semi-parametric frailty model has smaller AIC than its parametric counterparts. It is of interest to notice that estimates of

frailty variance are quite large for parametric lognormal and generalized gamma, and so are their standard errors. It is because we estimate the parameter of the frailty distribution (i.e., shape parameter for BS, gamma and inverse Gaussian and standard deviation of logarithm for lognormal), the estimates of frailty variance and its standard error are obtained by delta method. Small changes of estimate of parameter for lognormal distribution results in large change in estimate of its variance. The estimated CDF is presented in Figure 7. The black step curve is the non-parametric CDF obtained from the Kaplan–Meier estimates.

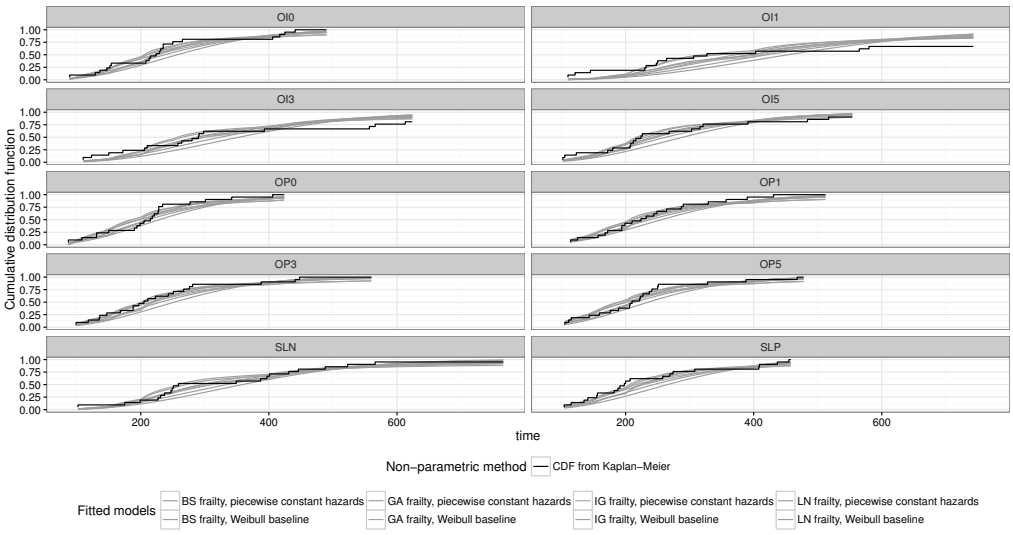


Figure 7: Fitted cumulative distribution functions.

We can see all the eight models fit the data well. To quantify the goodness-of-fit, we calculate the Kolmogorov–Smirnov distance (KSD) between the CDF of fitted models and the non-parametric CDF, presented in Table 5. It is defined as $D = \sup|\hat{F}(t) - \hat{F}_{km}(t)|$. It is seen clearly that piecewise linear baseline is better than Weibull baseline for all the models considered. This is also seen in the maximized log-likelihood and AIC values in Tables 3 and 4. Overall, the fits as measured by KSD are all quite similar with those by AIC indicating BS and IG models to be better. We also should examine the residuals to check the error. Figure 8 presents the deviance residuals, which is defined as

$$D_{r_{ij}} = \operatorname{sign}(r_{ij}) \sqrt{-2\left[r_{ij} + \delta_{ij} \log(\delta_{ij} - r_{ij})\right]},$$

where $r_{ij} = \delta_{ij} + \log(\hat{S}(t_{ij}))$. It can be seen that the deviance residuals are randomly distributed along 0. The deviance residuals should follow a standard normal distribution. For checking this, the QQ plot and envelopes of the deviance residuals are presented in Figure 9. It seems that all the models satisfy the normality assumption and the semi-parametric models are slightly better than the

parametric ones. The right tail of semi-parametric gamma frailty model deviates from the straight line more than the others. Semi-parametric BS, lognormal and inverse Gaussian frailty models are quite similar. Overall, semi-parametric BS is seen to be quite a robust model for modeling these clustered failure time data.

Table 5: KSD between estimated CDF and non-parametric CDF.

Frailty	Baseline	OI0	OI1	OI3	OI5	OP0	OP1	OP3	OP5	SLN	SLP	Overall
BS	piecewise	0.15	0.22	0.22	0.12	0.19	0.11	0.07	0.17	0.16	0.12	0.22
GA	piecewise	0.19	0.21	0.18	0.15	0.23	0.14	0.11	0.20	0.19	0.12	0.23
IG	piecewise	0.13	0.18	0.21	0.12	0.13	0.07	0.10	0.13	0.11	0.20	0.21
LN	piecewise	0.12	0.19	0.20	0.10	0.12	0.06	0.09	0.11	0.09	0.19	0.20
BS	Weibull	0.13	0.25	0.23	0.12	0.18	0.11	0.12	0.15	0.11	0.14	0.25
GA	Weibull	0.16	0.19	0.19	0.12	0.23	0.15	0.10	0.18	0.12	0.12	0.23
IG	Weibull	0.24	0.25	0.22	0.23	0.24	0.14	0.21	0.25	0.23	0.28	0.28
LN	Weibull	0.17	0.22	0.21	0.15	0.18	0.10	0.15	0.18	0.18	0.20	0.22

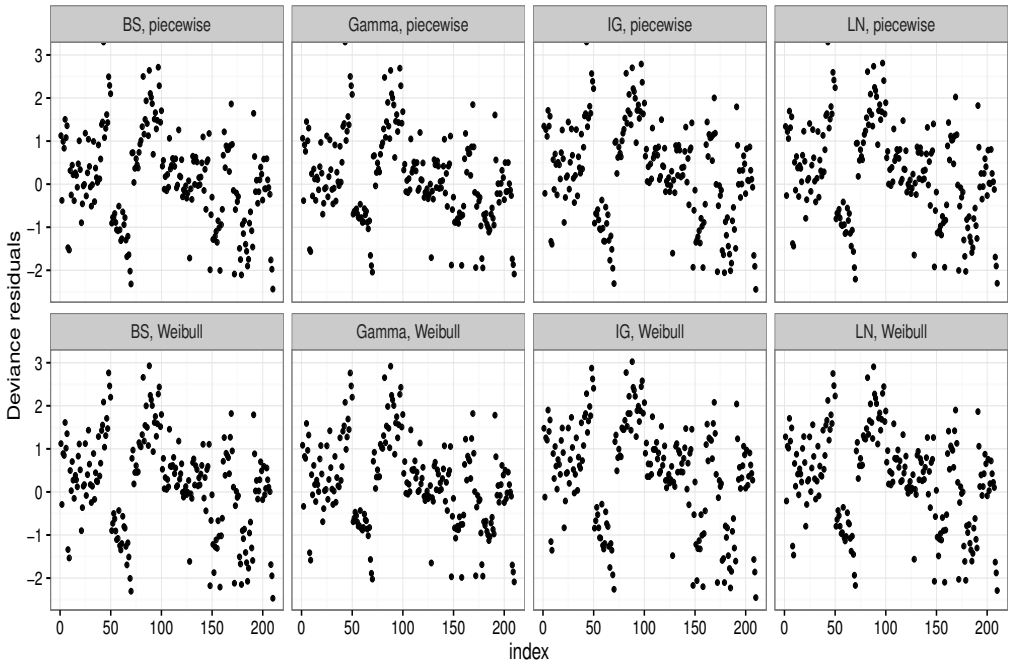


Figure 8: Deviance residuals.

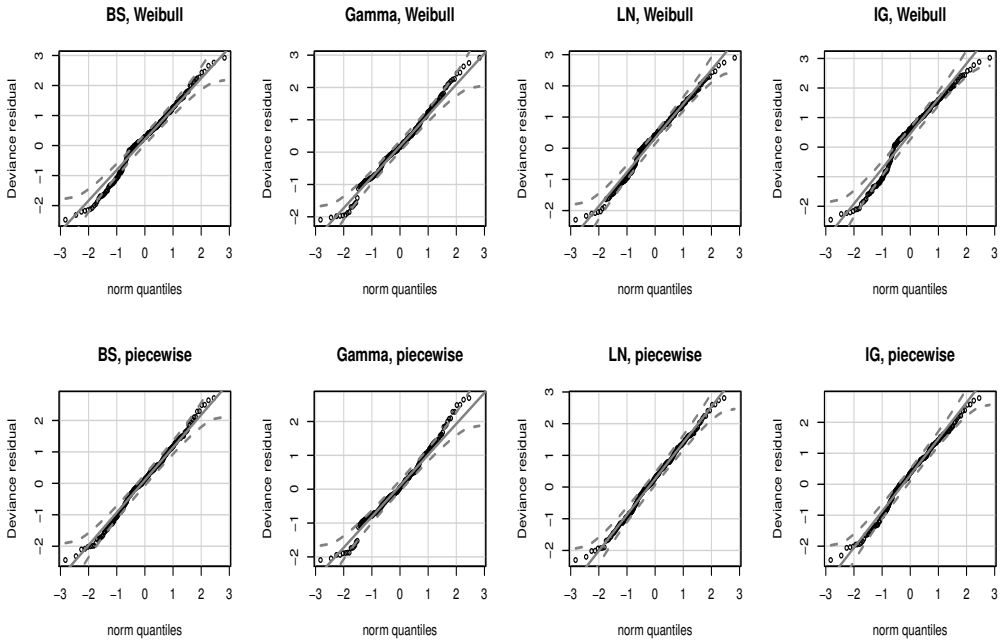


Figure 9: QQ plots for deviance residuals.

6. DISCUSSION AND CONCLUDING REMARKS

In this work, we have proposed a semi-parametric frailty model with BS frailty distribution. The non-parametric choice of baseline hazard function provides a robust and flexible way to model the data. The determination of MLEs becomes very difficult due to the intractable integrals present in the likelihood function. For this reason, Monte Carlo simulations are used to approximate the likelihood function upon exploiting the relationship between BS and standard normal distributions and then expressing those integrals as expectations of some functions of standard normal variables. From the simulation study carried out and the illustrative example analyzed, the semi-parametric BS frailty model is seen to be quite robust in estimating the covariate effects as well as the frailty variance. Interestingly, it is seen to be even better than the three-parameter generalized gamma frailty model though the latter has an extra shape parameter. It is of interest to mention that the work carried out here can be generalized in two different directions. The BS distribution can be generalized by assuming that the variable Z in (2.2) follows a standard elliptically symmetric distribution, including power exponential, Laplace, Student t and logistic distributions. Such a generalized Birnbaum–Saunders (GBS) distribution (see [14]) could be assumed for the frailty term y_i in (1.1) and then the resulting GBS frailty model could be studied in detail. Next, we could allow for the possibility of a cure of patients within the context of BS frailty model and develop the corresponding analysis. Work is currently under progress on these problems and we hope to report these findings in a future paper.

APPENDIX A — FIRST- AND SECOND-ORDER DERIVATIVES OF THE LOG-LIKELIHOOD FUNCTION

The first- and second-order derivatives of the log-likelihood function with respect to $\alpha, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are as follows:

$$
\begin{aligned}
\frac{\partial l}{\partial \alpha} &= \sum_{i=1}^{n} \frac{1}{I_i} \frac{\partial I_i}{\partial \alpha}, \\
\frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left[\sum_{j=1}^{m_i} \delta_{ij} \boldsymbol{x_{ij}} + \frac{1}{I_i} \frac{\partial I_i}{\partial \boldsymbol{\beta}} \right], \\
\frac{\partial l}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n} \left[\sum_{j=1}^{m_i} \frac{\delta_{ij}}{h_0(t_{ij})} \frac{dh_0(t_{ij})}{d\boldsymbol{\gamma}} + \frac{1}{I_i} \frac{\partial I_i}{\partial \boldsymbol{\gamma}} \right]; \\
\frac{\partial^2 l}{\partial \alpha^2} &= \sum_{i=1}^{n} \left[-\frac{1}{I_i^2} \left(\frac{\partial I_i}{\partial \alpha} \right)^2 + \frac{1}{I_i} \frac{\partial^2 I_i}{\partial \alpha^2} \right], \\
\frac{\partial^2 l}{\partial \alpha \partial \boldsymbol{\beta}^T} &= \sum_{i=1}^{n} \left[-\frac{1}{I_i^2} \frac{\partial I_i}{\partial \alpha} \left(\frac{\partial I_i}{\partial \boldsymbol{\beta}} \right)^T + \frac{1}{I_i} \frac{\partial^2 I_i}{\partial \alpha \partial \boldsymbol{\beta}^T} \right], \\
\frac{\partial^2 l}{\partial \alpha \partial \boldsymbol{\gamma}^T} &= \sum_{i=1}^{n} \left[-\frac{1}{I_i^2} \frac{\partial I_i}{\partial \alpha} \left(\frac{\partial I_i}{\partial \boldsymbol{\gamma}} \right)^T + \frac{1}{I_i} \frac{\partial^2 I_i}{\partial \alpha \partial \boldsymbol{\gamma}^T} \right], \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \sum_{i=1}^{n} \left[-\frac{1}{I_i^2} \frac{\partial I_i}{\partial \boldsymbol{\beta}} \left(\frac{\partial I_i}{\partial \boldsymbol{\beta}} \right)^T + \frac{1}{I_i} \frac{\partial^2 I_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right], \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} &= \sum_{i=1}^{n} \left[-\frac{1}{I_i^2} \frac{\partial I_i}{\partial \boldsymbol{\beta}} \left(\frac{\partial I_i}{\partial \boldsymbol{\gamma}} \right)^T + \frac{1}{I_i} \frac{\partial^2 I_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} \right], \\
\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} &= \sum_{i=1}^{n} \sum_{j=1}^{m_i} -\frac{\delta_{ij}}{h_0(t_{ij})^2} \frac{dh_0(t_{ij})}{d\boldsymbol{\gamma}} \left(\frac{dh_0(t_{ij})}{d\boldsymbol{\gamma}} \right)^T \\
&\quad + \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{\delta_{ij}}{h_0(t_{ij})} \frac{d^2 h_0(t_{ij})}{d\boldsymbol{\gamma}\, d\boldsymbol{\gamma}^T} \\
&\quad + \sum_{i=1}^{n} \left[-\frac{1}{I_i^2} \frac{\partial I_i}{\partial \boldsymbol{\gamma}} \left(\frac{\partial I_i}{\partial \boldsymbol{\gamma}} \right)^T + \frac{1}{I_i} \frac{\partial^2 I_i}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} \right],
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{\partial I_i}{\partial \alpha} &= \delta_{i\cdot} E_{1,i} - \left[\sum_{j=1}^{m_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}) \right] E_{2,i}, \\
\frac{\partial^2 I_i}{\partial \alpha^2} &= \delta_{i\cdot}(\delta_{i\cdot} - 1) E_{3,i} - 2\delta_{i\cdot} \left[\sum_{j=1}^{m_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}) \right] E_{4,i} + \delta_{i\cdot} E_{5,i} \\
&\quad + \left[\sum_{j=1}^{m_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}) \right]^2 E_{6,i} - \left[\sum_{j=1}^{m_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}) \right] E_{7,i},
\end{aligned}
$$

$$\begin{aligned}
\frac{\partial^2 I_i}{\partial\alpha\,\partial\boldsymbol{\beta}^T} &= \left\{\left[\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\right]E_{8,i} - (\delta_{i\cdot}+1)E_{2,i}\right\} \\
&\quad\times\left[\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\boldsymbol{x}_{ij}\right], \\
\frac{\partial I_i}{\partial\boldsymbol{\beta}} &= -E_{9,i}\left[\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\boldsymbol{x}_{ij}\right], \\
\frac{\partial I_i}{\partial\boldsymbol{\gamma}} &= -E_{9,i}\left[\sum_{j=1}^{m_i} \frac{dH_0(t_{ij})}{d\boldsymbol{\gamma}}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\right], \\
\frac{\partial^2 I_i}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^T} &= E_{10,i}\left[\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\boldsymbol{x_{ij}}\right]\left[\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\boldsymbol{x_{ij}}\right]^T \\
&\quad - E_{9,i}\left[\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\boldsymbol{x_{ij}}\,\boldsymbol{x_{ij}}^T\right], \\
\frac{\partial^2 I_i}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\gamma}^T} &= E_{10,i}\left[\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\boldsymbol{x_{ij}}\right]\left[\sum_{j=1}^{m_i} \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\frac{dH_0(t_{ij})}{d\boldsymbol{\gamma}}\right]^T \\
&\quad - E_{9,i}\left[\sum_{j=1}^{m_i} \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\boldsymbol{x_{ij}}\left(\frac{dH_0(t_{ij})}{d\boldsymbol{\gamma}}\right)^T\right], \\
\frac{\partial^2 I_i}{\partial\boldsymbol{\gamma}\,\partial\boldsymbol{\gamma}^T} &= E_{10,i}\left[\sum_{j=1}^{m_i} \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\frac{dH_0(t_{ij})}{d\boldsymbol{\gamma}}\right]\left[\sum_{j=1}^{m_i} \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\frac{dH_0(t_{ij})}{d\boldsymbol{\gamma}}\right]^T \\
&\quad - E_{9,i}\left[\sum_{j=1}^{m_i} \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\,\frac{d^2H_0(t_{ij};\boldsymbol{\gamma})}{d\boldsymbol{\gamma}\,d\boldsymbol{\gamma}^T}\right];
\end{aligned}$$

in the above expressions, the quantities $E_{l,i}$ $(l = 1, ..., 10)$ are given by

$$\begin{aligned}
E_{1,i} &= \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}-1}\exp\Big(-g(z_{(k)})\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big)\frac{dg(z_{(k)})}{d\alpha}, \\
E_{2,i} &= \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}}\exp\Big(-g(z_{(k)})\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big)\frac{dg(z_{(k)})}{d\alpha}, \\
E_{3,i} &= \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}-2}\exp\Big(-g(z_{(k)})\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big)\left(\frac{dg(z_{(k)})}{d\alpha}\right)^2, \\
E_{4,i} &= \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}-1}\exp\Big(-g(z_{(k)})\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big)\left(\frac{dg(z_{(k)})}{d\alpha}\right)^2, \\
E_{5,i} &= \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}-1}\exp\Big(-g(z_{(k)})\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big)\frac{d^2g(z_{(k)})}{d\alpha^2},
\end{aligned}$$

$$E_{6,i} = \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}} \exp\Big(-g(z_{(k)})\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big)\left(\frac{dg(z_{(k)})}{d\alpha}\right)^2,$$

$$E_{7,i} = \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}} \exp\Big(-g(z_{(k)})\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big)\frac{d^2g(z_{(k)})}{d\alpha^2},$$

$$E_{8,i} = \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}+1} \exp\Big(-g(z_{(k)})\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big)\frac{dg(z_{(k)})}{d\alpha},$$

$$E_{9,i} = \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}+1} \exp\Big(-g(z_{(k)})\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big),$$

$$E_{10,i} = \frac{1}{N}\sum_{k=1}^{N} g(z_{(k)})^{\delta_{i\cdot}+2} \exp\Big(-g(z_{(k)})\sum_{j=1}^{m_i} H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\Big).$$

ACKNOWLEDGMENTS

The authors express their sincere thanks to the guest editor, Professor Sat Gupta, for extending an invitation and to the anonymous reviewers for their useful comments and suggestions on an earlier version of this manuscript which led to this improved version.

REFERENCES

[1] Balakrishnan, N.; Leiva, V. and López, J. (2007). Acceptance sampling plans from truncated life tests based on the generalized Birnbaum–Saunders distribution, *Communications in Statistics–Simulation and Computation*, **36**(3), 643–656.

[2] Balakrishnan, N. and Peng, Y. (2006). Generalized gamma frailty model, *Statistics in Medicine*, **25**(16), 2797–2816.

[3] Birnbaum, Z.W. and Saunders, S.C. (1969). A new family of life distributions, *Journal of Applied Probability*, **55**(2), 319–327.

[4] Chang, D.S. and Tang, L.C. (1993). Reliability bounds and critical time for the Birnbaum–Saunders distribution, *IEEE Transactions on Reliability*, **42**(3), 464–469.

[5] Cox, D.R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society, Series B*, **34**(2), 187–220.

[6] Danahy, D.T.; Burwell, D.T.; Aronow, W.S. and Prakash, R. (1977). Sustained hemodynamic and antianginal effect of high dose oral isosorbide dinitrate, *Circulation*, **55**(2), 381–387.

[7] Desmond, A. (1985). Stochastic models of failure in random environments, *The Canadian Journal of Statistics*, **13**(3), 171–183.

[8] Dupuis, D.J. and Mills, J.E. (1998). Robust estimation of the Birnbaum–Saunders distribution, *IEEE Transactions on Reliability*, **47**(1), 88–95.

[9] From, S.G. and Li, L. (2006). Estimation of the parameters of the Birnbaum–Saunders distribution, *Communications in Statistics — Theory and Methods*, **35**(12), 2157–2169.

[10] Hougaard, P. (1986a). A class of multivariate failure time distributions, *Biometrika*, **73**(3), 671–678.

[11] Hougaard, P. (1986b). Survival models for heterogeneous populations derived from stable distributions, *Biometrika*, **73**(2), 387–396.

[12] Hougaard, P. (2012). *Analysis of Multivariate Survival Data*, Springer, New York.

[13] Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm, *Biometrics*, **48**(3), 795–806.

[14] Leiva, V.; Riquelme, M.; Balakrishnan, N. and Sanhueza, A. (2008). Lifetime analysis based on the generalized Birnbaum–Saunders distribution, *Computational Statistics & Data Analysis*, **52**(4), 2079–2097.

[15] Lemonte, A.J.; Cribari-Neto, F. and Vasconcellos, K.L. (2007). Improved statistical inference for the two-parameter Birnbaum–Saunders distribution, *Computational Statistics & Data Analysis*, **51**(9), 4656–4681.

[16] McGilchrist, C. and Aisbett, C. (1991). Regression with frailty in survival analysis, *Biometrics*, **47**(2), 461–466.

[17] Ng, H.K.T.; Kundu, D. and Balakrishnan, N. (2003). Modified moment estimation for the two-parameter Birnbaum–Saunders distribution, *Computational Statistics & Data Analysis*, **43**(3), 283–298.

[18] Ng, H.K.T.; Kundu, D. and Balakrishnan, N. (2006). Point and interval estimation for the two-parameter Birnbaum–Saunders distribution based on type-II censored samples, *Computational Statistics & Data Analysis*, **50**(11), 3222–3242.

[19] Rieck, J.R. (1999). A moment-generating function with application to the Birnbaum–Saunders distribution, *Communications in Statistics — Theory and Methods*, **28**(9), 2213–2222.

[20] Vaupel, J.W.; Manton, K.G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, **16**(3), 439–454.

REVSTAT – Statistical Journal
Volume 16, Number 2, April 2018, 257–278

ASSOCIATION MEASURES IN THE BIVARIATE CORRELATED FRAILTY MODEL

Author: Ramesh C. Gupta
– Department of Mathematics and Statistics, University of Maine, Orono, Maine 04469-5752, USA
ramesh_gupta@umit.maine.edu

Received: March 2017 Revised: October 2017 Accepted: October 2017

Abstract:

- This paper deals with a general bivariate correlated frailty model. This includes the multiplicative as well as the additive frailty effect. The association parameter is studied for the shared as well as the general correlated model. The results for the gamma, the inverse Gaussian and the stable frailty models are derived.

Key-Words:

- *survival function; failure rate; Clayton's association measure; gamma distribution; inverse Gaussian distribution; positive stable distribution.*

AMS Subject Classification:

- 62H20, 62N99.

1. INTRODUCTION

Cox (1972) proportional hazard model (PHM) is commonly used to model survival data as a function of the covariates. Sometimes the observed source of variation in the explanatory variables fail to account for the true differences in risk. That is,in addition, there are other important but omitted unobserved variables present. These unobserved random effects are modeled by introducing a frailty variable Z. More precisely we assume that (T, Z) is a pair of non-negative random variables such that for each z in the support of the distribution of Z, the conditional distribution of T given $Z = z$ is absolutely continuous with hazard rate $\lambda(t|z)$ given by

$$\lambda(t|z) = z\lambda_0(t), \quad t > 0, \tag{1.1}$$

where $\lambda_0(t)$ is the base line hazard rate independent of z. It will be helpful to think of T as the age at death and $\lambda(t|z)$ as the hazard rate at age t for a person with frailty Z, see Vaupel *et al.* (1979).

The model (1.1) states that the hazard rate of an individual is the product of the specific quantity z and the base line hazard $\lambda_0(t)$ describing the age.

In addition to introducing the unobserved random effects in a multiplicative manner, various other forms have been studied in the literature in the context of random effect models. More recently, there has been an interest in studying additive frailty models. Tomazalla *et al.* (2006) have analyzed recurrent event data considering a homogeneous Poisson process with additive frailty intensity. Silva and Amaral Turkman (2004) have considered Bayesian analysis of an additive survival model with frailty. Yin and Ibrahim (2005) presented a class of Bayesian shared Gamma frailty models with multivariate failure time data.

In this paper, we shall study a very general frailty model where the conditional failure rate $\lambda(t|z) = \lambda(t, z)$, is an appropriate general function of t and z. Obviously, the multiplicative (proportional hazards) as well as the additive model can be studied under this umbrella.

A basic problem in a frailty model is the modeling of the probability distribution of Z. The choice of the frailty distribution strongly affects the estimate of the base line hazard as well as that of the conditional probabilities, see Hougaard (1984, 19991, 1995, 2000), Heckman and Singer (1984) and Agresti *et al.* (2004). Agresti *et al.* (2004) have demonstrated that a considerable loss of efficiency can result from assuming a parametric distribution for a random effect that is substantially different from that of the true population.These authors observed that the misspecification of random effect has the potential for a serious drop of efficiency in the prediction of random effects and the estimation of other parameters. In the absence of a theoretical basis for selecting the distribution of

frailty, the choice of the distribution of Z is often made on the basis of mathematical tractability and the nice properties of the resulting distributions. For this reason, frailty distributions having a tractable Laplace transform are natural choices. The gamma distribution, the inverse Gaussian distribution and the family of stable distributions are popular choices for modeling the distribution of Z. Some researchers propose nonparametric modeling of the frailty distribution, see Heckman and Singer (1984) and Anderson *et al.* (1992).

Hougaard (2000) provides some guidelines for choosing an appropriate frailty distribution.The comparison is made in three directions:

(**1**) Theoretical comparison describing the nice properties of the frailty distribution. For example, the gamma distribution and the inverse Gaussian distribution are easibly tractable.

(**2**) Comparison of fit: The fit and the flexibility of the models are important factors in comparison. The stable frailty distribution implies high early dependence, whereas the gamma frailty model describes high late dependence.

(**3**) Various measures of dependence: The measures of dependence depend on the frailty distribution. The expressions for various dependence measures depend on the frailty distribution. For some frailty distributions, it is simple to evaluate these measures.

For more discussion, see Hougaard (2000).

Since different level distributions of frailty give rise to different population level distribution for analyzing survival data, it is appropriate to investigate how the comparative effect of two frailties translates into the comparative effect on the survival distribution. The stochastic orderings on various characteristics of the model can be studied by using the general results contained in Gupta and Gupta (2009, 2010). Also see Gupta and Kirmani (2006).

The aim of this paper is to study a general bivariate correlated frailty model and the association measure due to Clayton (1978). The bivariate correlated model and its derivatives have been studied in the literature in the context of twin's survival, see for example Yashin and Iachine (1995a, 1995b) and Yashin *et al.* (1995). The idea of using the shared relative risk in bivariate survival models was first discussed by Clayton (1978) who suggested an approach to the analysis of association between two survival times based on the limiting properties of certain contingency tables. Later this approach was followed by Oakes (1989) in the proportional hazards shared frailty model. He introduced the notion of the local association measure which characterizes the limiting behaviour of the odds ratio statistics for the dependent life spans. The properties of this measure were studied by Anderson *et al.* (1992).

We obtain a general expression for the population level survival function. The proportional hazards as well as the additive hazards case is studied. General expressions for the Clayton's(1978) association measure are obtained.The results are illustrated for the gamma frailty model and the inverse Gaussian frailty model.

The organization of this paper is as follows: Section 2 contains the general bivariate correlated frailty model and an expression for the population level survival function. Explicit expressions are obtained for the bivariate gamma correlated model. The Clayton's association measure is studied in Section 3. Results are derived for the multiplicative as well as the additive frailty models. Several examples are provided. It also contains the results for the shared frailty model. Section 4 contains some practical examples from the literature. Finally, some conclusions and comments are provided in Section 5.

2. BIVARIATE CORRELATED FRAILTY MODEL

Let T_i and Z_i, $i = 1, 2$ be the life spans and frailty variables for the two related individuals with dependent individual hazards $\mu_i(x_i, Z_i)$, $i = 1, 2$. The functional form of $\mu_i(x_i, Z_i)$ is assumed to be the same for both individuals. We assume that the life spans T_1 and T_2 are conditionally independent given Z_1 and Z_2. Also the joint, conditional and marginal distributions are absolutely continuous.

Then the joint conditional survival function of T_1 and T_2 is given by

$$S(x_1, x_2|z_1, z_2) = \exp\big\{-\big(H_1(x_1, z_1) + H_2(x_2, z_2)\big)\big\},$$

where

$$H_i(x_i, z_i) = \int_0^{x_i} \mu_i(u_i, z_i) du_i, \quad i = 1, 2.$$

The unconditional survival function is given by

$$S(x_1, x_2) = \iint \exp\big\{-\big(H_1(x_1, z_1) + H_2(x_2, z_2)\big)\big\}\, g(z_1, z_2)\, dz_1\, dz_2,$$

where $g(z_1, z_2)$ is the joint probability density function (pdf) of (Z_1, Z_2).

This gives

$$\begin{aligned} S_i(x_1, x_2) &= \frac{\partial}{\partial x_i} S(x_1, x_2) \\ &= -E_{Z_1,Z_2}\big[\mu_i(x_i, Z_i) \exp\big\{-\big(H_1(x_1, Z_1) + H_2(x_2, Z_2)\big)\big\}\big], \quad i = 1, 2, \end{aligned}$$

and

$$\begin{aligned} f(x_1, x_2) &= \frac{\partial^2}{\partial x_1 \partial x_2} S(x_1, x_2) \\ &= E_{Z_1,Z_2}\big[\mu_1(x_1, Z_1)\, \mu_2(x_2, Z_2) \exp\big\{-\big(H_1(x_1, Z_1) + H_2(x_2, Z_2)\big)\big\}\big]. \end{aligned}$$

Thus

$$\begin{aligned}\frac{f(x_1,x_2)}{S(x_1,x_2)} &= \frac{1}{S(x_1,x_2)}\iint \mu_1(x_1,z_1)\,\mu_2(x_2,z_2)\\ &\quad\times \exp\big\{-\big(H_1(x_1,z_1)+H_2(x_2,z_2)\big)\big\}\,g(z_1,z_2)\,dz_1\,dz_2\\ &= \iint \mu_1(x_1,z_1)\,\mu_2(x_2,z_2)\,g\big(z_1,z_2\,|\,T_1>x_1,T_2>x_2\big)\,dz_1\,dz_2 \qquad (2.1)\\ &= \rho_{\mu_1,\mu_2}(x_1,x_2)\,\sigma_{\mu_1}(x_1,x_2)\,\sigma_{\mu_2}(x_1,x_2)+\overline{\mu_1}(x_1,x_2)\,\overline{\mu_2}(x_1,x_2),\end{aligned}$$

where

$$\overline{\mu_i}(x_1,x_2) = E\big[\mu_i(x_i,Z_i)\,|\,T_1>x_1,T_2>x_2\big],\quad i=1,2,$$

$\rho(.,.)$ is the conditional correlation coefficient and $\sigma_{\mu_{ii}}|\,T_1>x_1,\,T_2>x_2,\,i=1,2$ is the conditional standard deviation.

Also

$$g\big(z_1,z_2\,|\,T_1>x_1,T_2>x_2\big) = \frac{\exp\big\{-\big(H_1(x_1,z_1)+H_2(x_2,z_2)\big)\big\}}{S(x_1,x_2)}\;g(z_1,z_2)$$

is the conditional pdf of Z_1,Z_2 given $T_1>x_1,T_2>x_2$.

The hazard components are given by

$$\begin{aligned}h_i(x_1,x_2) &= -\frac{\partial}{\partial x_i}\ln S(x_1,x_2)\\ &= -\iint \mu_i(x_i,z_i)\,g\big(z_1,z_2\,|\,T_1>x_1,T_2>x_2\big)\,dz_1\,dz_2\\ &= -E\big[\mu_i(x_i,Z_i)\,|\,T_1>x_1,T_2>x_2\big] = -\overline{\mu_i}(x_1,x_2),\quad i=1,2.\end{aligned}$$

Note that the expectations are taken with respect to the conditional distribution of the joint distribution of the frailty given $T_1>x_1,T_2>x_2$.

Define

$$\begin{aligned}\phi(x_1,x_2) &= \frac{\partial^2}{\partial x_1\partial x_2}\ln S(x_1,x_2)\\ &= \frac{f(x_1,x_2)}{S(x_1,x_2)} - h_1(x_1,x_2)\,h_2(x_1,x_2)\\ &= E\big[\mu_1(x_1,Z_1)\,\mu_2(x_2,Z_2)\,|\,T_1>x_1,T_2>x_2\big]\\ &\quad - E\big[\mu_1(x_1,Z_1)\,|\,T_1>x_1,T_2>x_2)\big]\big[E\big[\mu_2(x_2,Z_2)\,|\,T_1>x_1,T_2>x_2\big]\big]\\ &= Cov\big[\mu_1(x_1,Z_1),\mu_2(x_2,Z_2)\,|\,T_1>x_1,T_2>x_2\big]\\ &= \rho\big[\mu_1(x_1,Z_1),\mu(x_2,Z_2)\,|\,T_1>x_1,T_2>x_2\big]\\ &\quad\times\big[\sigma_{\mu_1(x_1,Z_1)}\,|\,T_1>x_1,T_2>x_2\big]\big[\sigma_{\mu_2(x_2,Z_2)}\,|\,T_1>x_1,T_2>x_2\big].\end{aligned}$$

Let

$$A(x_1,x_2) = \int_0^{x_2}\int_0^{x_1}\phi(u_1,u_2)\,du_1\,du_2.$$

Thus

$$\ln S(x_1, x_2) \;=\; \int_0^{x_2}\!\int_0^{x_1} \phi(u_1, u_2)\, du_1\, du_2 \;-\; \int_0^{x_1} \psi_1(u)\, du \;-\; \int_0^{x_2} \psi_2(u)\, du,$$

for some appropriate functions $\psi_1(.)$ and $\psi_2(.)$.

Finally,

$$\begin{aligned} S(x_1, x_2) \;&=\; \exp\Big\{ \int_0^{x_2}\!\int_0^{x_1} \phi(u_1, u_2)\, du_1\, du_2 \;-\; \int_0^{x_1} \psi_1(u)\, du \;-\; \int_0^{x_2} \psi_2(u)\, du \Big\} \\ &=\; S_1(x_1)\, S_2(x_2)\, \exp\{A(x_1, x_2)\}, \end{aligned}$$

where

$$S_i(x_i) \;=\; \exp\Big\{ -\int_0^{x_i} \psi_i(u)\, du \Big\}, \quad i = 1, 2.$$

We now present a bivariate gamma correlated frailty model.

2.1. Bivariate Gamma Correlated Frailty Model

Suppose Y_0, Y_1 and Y_2 are independent random variables and $Z_1 = Y_0 + Y_1$, $Z_2 = Y_0 + Y_2$. Then Z_1 and Z_2 are correlated since they contain the common part Y_0. This constitutes one of the ways of constructing bivariate distributions, see Marshall and Olkin (1988). Let

$$S(x_1, x_2 | z_1, z_2) \;=\; \exp\big\{ -\big(H_1(x_1) z_1 + H_2(x_2) z_2 \big) \big\}, \tag{2.2}$$

i.e., given Z_1 and Z_2, the life spans T_1 and T_2 are independent. This is the proportional hazards bivariate correlated model. The unconditional distribution is given by

$$\begin{aligned} S(x_1, x_2) \;=\; &\iiint \exp\big\{ -\big((y_0 + y_1)\, H_1(x_1) + (y_0 + y_2)\, H_2(x_2) \big) \big\} \\ &\times\, g_0(y_0)\, g_1(y_1)\, g_2(y_2)\, dy_0\, dy_1\, dy_2, \end{aligned} \tag{2.3}$$

where $g_0(.)$, $g_1(.)$ and $g_2(.)$ are the *pdf's* of Y_0, Y_1 and Y_2. Denoting by $L_{Y_0}(.)$, $L_{Y_1}(.)$ and $L_{Y_2}(.)$ the Laplace transform of Y_0, Y_1 and Y_2, it can be seen that

$$S(x_1, x_2) \;=\; L_{Y_0}\big[H_1(x_1) + H_2(x_2) \big]\, L_{Y_1}[H_1(x_1)]\, L_{Y_2}[H_2(x_2)]. \tag{2.4}$$

We shall now derive the correlated frailty model of Yashin *et al.* (1995); see also Korsgaard and Anderson (1998).

Let Y_0, Y_1, Y_2 have independent gamma distribution with parameters (α_0, β_0), (α_1, β_1) and (α_2, β_2) having *pdf's*

$$g_i(y_i) = \frac{1}{\beta_i^{\alpha_i} \Gamma(\alpha_i)}\; e^{-y_i/\beta_i}\, y_i^{\alpha_i - 1}, \qquad y_i > 0, \quad i = 0, 1, 2. \tag{2.5}$$

To ensure that Z_1 and Z_2 are gamma distributed, we make the assumption (on the scale parameters) that $\beta_0 = \beta_1 = \beta_2 = \beta$ (say). Note that this assumption is not a restriction for population of unrelated individuals since gamma distributed variables $Z_i, i = 1, 2$ can be decomposed this way. Thus

$$E(Z_1) = (\alpha_0 + \alpha_1)\beta, \qquad E(Z_2) = (\alpha_0 + \alpha_2)\beta,$$
$$Var(Z_1) = (\alpha_0 + \alpha_1)\beta^2, \qquad Var(Z_2) = (\alpha_0 + \alpha_2)\beta^2.$$

We now assume that Z_1 and Z_2 have the same gamma distribution. To do this, we assume that $\alpha_1 = \alpha_2 = \alpha$ (say). This condition is relevant in twin studies when there is no reason to assume different distributions of frailty for the twins.

The correlation coefficient between Z_1 and Z_2 is

$$\rho_Z = \frac{Var(Y_0)}{\sqrt{Var(Z_1)Var(Z_2)}} = \frac{\alpha_0}{\alpha_0 + \alpha}.$$

This implies that $\alpha_0 = \alpha\rho_Z/(1 - \rho_Z)$.

We now use the standard assumption that the mean frailty of the individuals is 1. This condition is typical for proportional hazards models which do not contain a frailty term, but covariates. This will imply that $Var(Z_1) = Var(Z_2) = \beta = \sigma_Z^2$ (say) and hence $\alpha_0 = \rho_Z/\sigma_Z^2$. Note that the formulated assumptions significantly restrict the class of frailty models which we propose here. However, this class is still wide enough to include individual frailty models and shared frailty models with gamma distributed random effects as particular cases.

Noting that $L_{Y_0}(t) = (1 + \beta t)^{-\alpha_0}$, $L_{Y_1}(t) = L_{Y_2}(t) = (1 + \beta t)^{-\alpha}$, it can be verified that

$$\begin{aligned} S(x_1, x_2) \;=\; & \big[1 + \sigma_Z^2\big(H_1(x_1) + H_2(x_2)\big)\big]^{-\rho_Z/\sigma_Z^2} \\ & \times \big[\big(1 + \sigma_Z^2(H_1(x_1))\big)\big(1 + \sigma_z^2(H_2(x_2))\big)\big]^{-(1-\rho_Z)/\sigma_Z^2}. \end{aligned} \tag{2.6}$$

Shared Frailty Model

In the shared frailty model, the two shared components are identical and, therefore, $\rho_Z = 1$. The survival function is given by

$$S(x_1, x_2) \;=\; \big[1 + \sigma_Z^2\big(H_1(x_1) + H_2(x_2)\big)\big]^{-(1/\sigma_Z^2)}. \tag{2.7}$$

Remark 2.1. Recently Hanagal and Dabade (2015) have considered four shared frailty models. These models have been illustrated with real life bivariate survival data related to kidney infection.

3. CLAYTON'S ASSOCIATION MEASURE

In the context of bivariate survival models induced by frailties, Oakes (1989) studied the following association measure

$$\theta(x_1, x_2) = \frac{SS_{12}}{S_1 S_2},$$

where $S = S(x_1, x_2)$ is the survival function, $S_{12} = \partial^2 S(x_1, x_2)/\partial x_1 \partial x_2$, $S_1 = \frac{\partial}{\partial x_1} S(x_1, x_2)$ and $S_2 = \frac{\partial}{\partial x_2} S(x_1, x_2)$; see also Clayton (1978).

Clayton (1978) presented the above association measure, deriving from the Cox model, in a study of the association between the life spans of fathers and their sons.

It can be easily seen that

$$\theta(x_1, x_2) = \frac{r(x_1 | T_2 = x_2)}{h_1(x_1, x_2)}.$$

The numerator is the hazard rate for sons at time x_1 given that their fathers died at x_2. The denominator is the hazard rate for sons at time x_1 given that their fathers live past x_2. Also

$$r(x_1 | T_2 = x_2) = -S_{12}/S_2 \qquad \text{and} \qquad h_1(x_1, x_2) = -S_1/S.$$

For the bivariate frailty model considered before, we have from (2.1)

$$\frac{f(x_1, x_2)}{S(x_1, x_2)} = \rho_{\mu_1,\mu_2}(x_1,x_2)\, \sigma_{\mu_1}(x_1,x_2)\, \sigma_{\mu_2}(x_1, x_2) + \overline{\mu_1}(x_1, x_2)\, \overline{\mu_2}(x_1, x_2),$$

and

$$\frac{S_1(x_1, x_2)}{S(x_1, x_2)} \frac{S_2(x_1, x_2)}{S(x_1, x_2)} = \overline{\mu_1}(x_1, x_2)\, \overline{\mu_2}(x_1, x_2).$$

Thus

$$\begin{aligned} \theta(x_1, x_2) &= 1 + \frac{\sigma_{\mu_1}(x_1,x_2)\, \sigma_{\mu_2}(x_1, x_2)}{\overline{\mu_1}(x_1, x_2)\, \overline{\mu_2}(x_1, x_2)}\, \rho_{\mu_1,\mu_2}(x_1,x_2) \\ &= 1 + [CV_{\mu_1}(x_1, x_2)]\, [CV_{\mu_2}(x_1, x_2)]\, \rho_{\mu_1,\mu_2}(x_1,x_2), \end{aligned} \tag{3.1}$$

where $CV_{\mu_i}(x_1, x_2)$ is the coefficient of variation, $i = 1, 2$.

Note that all expectations are taken with respect to the conditional distribution of (Z_1, Z_2) given $T_1 > x_1, T_2 > x_2$.

It is, therefore, clear that

$$\begin{aligned} \theta(x_1, x_2) > 1, &\quad \text{if } \rho_{\mu_1,\mu_2}(x_1,x_2) > 0 \\ < 1, &\quad \text{if } \rho_{\mu_1,\mu_2}(x_1,x_2) < 0 \\ = 1, &\quad \text{if } \rho_{\mu_1,\mu_2}(x_1,x_2) = 0. \end{aligned}$$

It is also clear that

$$\begin{aligned} \theta(x_1, x_2) > 1, &\quad \text{if } \phi(x_1,x_2) > 0 \\ < 1, &\quad \text{if } \phi(x_1, x_2) < 0 \\ = 1, &\quad \text{if } \phi(x_1,x_2) = 0. \end{aligned}$$

3.1. Proportional Hazards Bivariate Correlated Frailty Model

In this case

$$\begin{aligned} \mu_1(x_1, Z_1) &= Z_1 \mu_1(x_1), \\ \mu_2(x_2, Z_2) &= Z_2 \mu_2(x_2). \end{aligned}$$

It can be verified that

$$\begin{aligned} \rho_{\mu_1,\mu_2}(x_1,x_2) &= \frac{Cov\big(\mu_1(x_1, Z_1), \mu_2(x_2, Z_2)\big)}{\sqrt{Var(\mu_1(x_1, Z_1))\, Var(\mu_2(x_2, Z_2))}} \\ &= \frac{\mu_1(x_1)\, \mu_2(x_2)\, \rho_{Z_1,Z_2}(x_1, x_2)\, \sigma_{Z_1}(x_1, x_2)\, \sigma_{Z_2}(x_1, x_2)}{\mu_1(x_1)\, \mu_2(x_2)\, \sigma_{Z_1}(x_1, x_2)\, \sigma_{Z_2}(x_1, x_2)} \\ &= \rho_{Z_1,Z_2}(x_1, x_2). \end{aligned}$$

Also

$$CV_{\mu_i}(x_1, x_2) = CV_{Z_i}(x_1, x_2), \quad i = 1, 2.$$

Hence

$$\theta(x_1, x_2) = 1 + \rho_{Z_1,Z_2}(x_1, x_2)\, CV_{Z_1}(x_1, x_2)\, CV_{Z_2}(x_1, x_2). \tag{3.2}$$

Shared Bivariate Frailty Model

In this case $Z_1 = Z_2 = Z$ (say) and $\rho_{Z_1,Z_2}(x_1, x_2) = 1$, giving

$$\theta(x_1, x_2) = 1 + CV_Z^2(x_1, x_2).$$

We now try to give an explicit expression for $\theta(x_1, x_2)$.

The conditional survival function of T_1 and T_2 given $Z = z$ is

$$S\big(x_1, x_2|Z=z\big) \;=\; \exp\big\{-z\big(H_1(x_1) + H_2(x_2)\big)\big\}.$$

The unconditional survival function is given by

$$\begin{aligned} S(x_1, x_2) \;&=\; \int_0^\infty \exp\big\{-z\big(H_1(x_1) + H_2(x_2)\big)\big\}\, g(z)\, dz \\ &=\; L_Z\big(H_1(x_1) + H_2(x_2)\big), \end{aligned}$$

where $L_Z(.)$ is the Laplace transform of Z.

Thus, the conditional density of Z given $T_1 > x_1, T_2 > x_2$ is given by

$$g\big(z|\, T_1 > x_1, T_2 > x_2\big) \;=\; \frac{\exp\big\{-z\big(H_1(x_1) + H_2(x_2)\big)\big\}}{L_Z\big(H_1(x_1) + H_2(x_2)\big)}\; g(z).$$

It can be verified that

$$E\big[Z\,|\,T_1 > x_1, T_2 > x_2\big] = \frac{-L_Z^{'}\big(H_1(x_1) + H_2(x_2)\big)}{L_Z\big(H_1(x_1) + H_2(x_2)\big)}$$

and

$$E\big[Z^2\,|\,T_1 > x_1, T_2 > x_2\big] = \frac{L_Z^{''}\big(H_1(x_1) + H_2(x_2)\big)}{L_Z\big(H_1(x_1) + H_2(x_2)\big)}.$$

Hence

$$Var\big[Z\,|\,T_1 > x_1, T_2 > x_2\big] = \left[\frac{L_Z^{''}\big(H_1(x_1) + H_2(x_2)\big)}{L_Z\big(H_1(x_1) + H_2(x_2)\big)}\right] - \left[\frac{L_Z^{'}\big(H_1(x_1) + H_2(x_2)\big)}{L_Z\big(H_1(x_1) + H_2(x_2)\big)}\right]^2.$$

Using the above expressions, one can obtain $\theta(x_1, x_2)$.

We now consider some examples

Example 3.1. Z has a gamma distribution with probability density function (*pdf*)

$$g(z) = \frac{1}{\beta^\alpha \Gamma(\alpha)}\; e^{-z/\beta}\, z^{\alpha-1}, \qquad z > 0, \quad \alpha > 0, \quad \beta > 0. \tag{3.3}$$

The Laplace transform of Z is given by

$$L_Z(t) \;=\; \frac{1}{(1+\beta t)^\alpha}.$$

This gives

$$\frac{L_Z^{'}(t)}{L_Z(t)} = \frac{-\alpha\beta}{1+\beta t}$$

and

$$\frac{L_Z^{''}(t)}{L_Z(t)} = \frac{\alpha\beta^2(\alpha+1)}{(1+\beta t)^2}.$$

It can be easily verified that in this case

$$\theta(x_1, x_2) = 1 + \frac{1}{\alpha}. \tag{3.4}$$

Note that, in this case, $\theta(x_1, x_2)$ is independent of (x_1, x_2); see Hanagal (2011, page 83) Wienke (2010) and Duchateau and Janssen (2008).

Example 3.2. Z has an inverse Gaussian distribution with pdf

$$g(z) = \left(\frac{1}{2\pi a z^3}\right)^{1/2} \exp\left[-(bz-1)^2/(2az)\right], \qquad z, a, b > 0. \tag{3.5}$$

The Laplace transform of Z is given by

$$L_Z(t) = \exp\left[\frac{b}{a}\left(1 - \left(1 + \frac{2a}{b^2}\,t\right)^{1/2}\right)\right].$$

This gives

$$\frac{-L_Z^{'}(t)}{L_Z(t)} = \frac{1}{(b^2+2at)^{1/2}}$$

and

$$\frac{L_Z^{''}(t)}{L_Z(t)} = \frac{1 + a(b^2+2at)^{-1/2}}{b^2+2at}.$$

It can be verified that, in this case

$$\theta(x_1, x_2) = 1 + \frac{a^2}{\left[b^2 + 2a\left(H_1(x_1) + H_2(x_2)\right)\right]^{1/2}}. \tag{3.6}$$

Remark 3.1. Recently Hanagal and Bhanbure (2016) considered inverse Gaussian distribution as frailty distribution and three baseline distributions. They applied these three models to the analysis of kidney infection data.

Example 3.3. Z has a positive stable distribution with *pdf*

$$f_Z(z) = -\frac{1}{\pi z}\sum_{k=1}^{\infty}\frac{\Gamma(k\alpha+1)}{k!}\,(-z^{-\alpha})^k\sin(\alpha k\pi), \quad z>0,\ 0<\alpha<1, \tag{3.7}$$

see Duchateau and Janssen (2008) for more explanation and justification of this distribution as frailty distribution. Note that this density has infinite mean. Therefore, the variance is undetermined.

The Laplace transform of Z is given by

$$L_Z(t) = e^{-t^{\alpha}}, \quad 0<\alpha<1,$$

whose derivatives are given by

$$L_Z^{'}(t) = -\alpha t^{\alpha-1}L_Z(t)$$

and

$$L_Z^{''}(t) = L_Z(t)\left[\alpha^2 t^{2\alpha-2} - \alpha(\alpha-1)t^{\alpha-2}\right].$$

It can be verified that

$$\theta(x_{1,}x_2) = 1 + \frac{(1-\alpha)}{\alpha\left[H_1(x_1+H_2(x_2)\right]^{\alpha}}. \tag{3.8}$$

Bivariate Gamma Correlated Proportional Hazards Model

We follow the notations and assumptions given in section 2.1. The conditional survival function is given by

$$S\big(x_1,x_2|Z_1=z_1,Z_2=z_2\big) = S\big(x_1,x_2\,|\,z_1,z_2\big) = \exp\big\{-\big(H_1(x_1)z_1+H_2(x_2)z_2\big)\big\}.$$

Here Z_1 and Z_2 have been taken with the same marginal distribution, but correlated. This means that $Var(Z_1)=Var(Z_2)=\sigma_Z^2$ (say). Also the correlation coefficient between Z_1 and Z_2 will be denoted by ρ_Z.

We have

$$\rho_{Z_1,Z_2}(x_1,x_2) = \frac{Var_{Y_0}(x_1,x_2)}{\sigma_{Z_1}(x_1,x_2)\,\sigma_{Z_2}(x_1,x_2)}.$$

Under our assumptions

$$\alpha_0 = \rho_Z/\sigma_Z^2, \quad \alpha=(1-\rho_Z)/\sigma_Z^2, \quad \beta=\sigma_Z^2,$$

$$Var_{Y_0}(x_1,x_2) = \frac{\alpha_0\,\sigma_Z^4}{\left[1+\sigma_Z^2\big(H_1(x_1)+H_2(x_2)\big)\right]^2} = \frac{\rho_Z\,\sigma_Z^2}{\left[1+\sigma_Z^2\big(H_1(x_1)+H_2(x_2)\big)\right]^2},$$

$$Var_{Y_i}(x_1,x_2) = \frac{\alpha\sigma_Z^4}{\big[1+\sigma_Z^2 H_i(x_i)\big]^2} = \frac{(1-\rho_Z)\sigma_Z^2}{\big[1+\sigma_Z^2 H_i(x_i)\big]^2}, \qquad i=1,2.$$

These give

$$\begin{aligned} Var_{Z_i}(x_1,x_2) &= Var_{Y_o}(x_1,x_2) + Var_{Y_i}(x_1,x_2) \\ &= \frac{\rho_Z\sigma_Z^2\big[1+\sigma_Z^2 H_i(x_i)\big]^2 + (1-\rho_Z)\,\sigma_Z^2\big[1+\sigma_Z^2\big(H_1(x_1)+H_2(x_2)\big)\big]^2}{\big[1+\sigma_Z^2 H_i(x_i)\big]^2\big[1+\sigma_Z^2\big(H_1(x_1)+H_2(x_2)\big)\big]^2}, \\ & \qquad\qquad i=1,2. \end{aligned}$$

Thus

$$\begin{aligned} &\rho_{Z_1,Z_2}(x_1,x_2) = \\ &= \frac{Var_{Y_0}(x_1,x_2)}{\sigma_{Z_1}(x_1,x_2)\,\sigma_{Z_2}(x_1,x_2)} \\ &= \frac{\rho_Z\big[1+\sigma_Z^2 H_1(x_1)\big(1+\sigma_Z^2 H_2(x_2)\big)\big]}{\Big[\prod_{i=1}^{i=2}\Big\{\rho_Z\big[1+\sigma_Z^2 H_i(x_i)\big]^2 + (1-\rho_Z)\big[1+\sigma_Z^2\big(H_1(x_1)+H_2(x_2)\big)\big]^2\Big\}\Big]^{1/2}}, \end{aligned} \tag{3.9}$$

Now

$$\begin{aligned} &E\big(Z_i|\,T_1>x_1, T_2>x_2\big) = \\ &= E\big(Y_0|\,T_1>x_1, T_2>x_2\big) + E\big(Y_i|\,T_1>x_1, T_2>x_2\big) \\ &= \frac{\rho_Z\big[1+\sigma_Z^2 H_i(x_i)\big] + (1-\rho_Z)\big[1+\sigma_Z^2\big(H_1(x_1)+H_2(x_2)\big)\big]}{\big[1+\sigma_Z^2 H_i(x_i)\big]\big[1+\sigma_Z^2\big(H_1(x_1)+H_2(x_2)\big)\big]}, \quad i=1,2. \end{aligned} \tag{3.10}$$

Using the above expressions, the $\mathrm{CV}_{Z_i}(x_1,x_2)$ is given by

$$\begin{aligned} &\big[CV_{Z_i}(x_1,x_2)\big]^2 = \\ &= \frac{\rho_Z\sigma_Z^2\big[1+\sigma_Z^2 H_i(x_i)\big]^2 + (1-\rho_Z)\,\sigma_Z^2\big[1+\sigma_Z^2\big(H_1(x_1)+H_2(x_2)\big)\big]^2}{\big\{\rho_Z\big[1+\sigma_Z^2 H_i(x_i)\big] + (1-\rho_Z)\big[1+\sigma_Z^2\big(H_1(x_1)+H_2(x_2)\big)\big]\big\}^2}, \\ & \qquad\qquad i=1,2. \end{aligned} \tag{3.11}$$

Using the expressions of $\rho_{Z_1,Z_2}(x_1,x_2)$, $CV_{Z_1}(x_1,x_2)$ and $CV_{Z_2}(x_1,x_2)$, $\theta(x_1,x_2)$ can be obtained.

Remark 3.2. Eriksson and Scheike (2015) have mentioned a similar formula, in the competing risk set up, in a more complex form. See also Gorfine and Hsu (2011) where they provide a new class of frailty based competing risk model for clustered failure time data.

Shared Frailty Model

In this case $\rho_Z = 1$ and hence $\rho_{Z_1,Z_2}(x_1,x_2) = 1$ and the expression for $\theta(x_1,x_2)$ simplifies to

$$\theta(x_1,x_2) = 1+\sigma_Z^2.$$

3.2. Additive Bivariate Correlated Frailty Model

In this case

$$\mu_1(x_1, Z_1) = Z_1 + \mu_1(x_1),$$
$$\mu_2(x_2, Z_2) = Z_2 + \mu_2(x_2).$$

It can be verified that

$$\begin{aligned}\rho_{\mu_1,\mu_2}(x_1,x_2) &= \frac{Cov\big(\mu_1(x_1,Z_1),\mu_2(x_2,Z_2)\big)}{\sqrt{Var(\mu_1(x_1,Z_1))\,Var(\mu_2(x_2,Z_2))}}\\ &= \rho_{Z_1,Z_2}(x_1,x_2).\end{aligned}$$

Also

$$CV_{\mu_i}(x_1,x_2) = \frac{\sqrt{Var_{Z_i}(x_1,x_2)}}{\mu_i(x_i) + E\big(Z_i|\,T_1>x_1, T_2>x_2\big)}, \quad i=1,2.$$

Hence

$$\begin{aligned}\theta(x_1,x_2) \;=\; 1 &+ \rho_{Z_1,Z_2}(x_1,x_2)\,\frac{\sqrt{Var(Z_1|\,T_1>x_1,T_2>x_2)}}{\big(\mu_1(x_1)+E(Z_1|\,T_1>x_1,T_2>x_2)\big)}\\ &\times \frac{\sqrt{Var(Z_2|\,T_1>x_1,T_2>x_2)}}{\big(\mu_2(x_2)+E(Z_2|\,T_1>x_1,T_2>x_2)\big)}.\end{aligned} \tag{3.12}$$

Shared Additive Bivariate Frailty Model

In this case $Z_1 = Z_2 = Z$ (say) and $\rho_{Z_1,Z_2}(x_1,x_2) = 1$, giving

(3.13)
$$\begin{aligned}&\theta(x_1,x_2) \;=\\ &= 1 + \frac{Var\big(Z\,|\,T_1>x_1,T_2>x_2\big)}{\big[\mu_1(x_1)+E\big(Z|T_1>x_1,T_2>x_2\big)\big]\big[\mu_2(x_2)+E\big(Z|T_1>x_1,T_2>x_2\big)\big]}.\end{aligned}$$

We now try to give an explicit expression for $\theta(x_1,x_2)$.

The conditional survival function of T_1 and T_2 given $Z=z$ is

$$S(x_1,x_2|Z=z) \;=\; \exp\big\{-\big(\Lambda_1(x_1)+\Lambda_2(x_2)+z(x_1+x_2)\big)\big\}$$

where $\Lambda_1(x_1)$ and $\Lambda_2(x_2)$ are the integrated hazards.

The unconditional survival function is given by

$$\begin{aligned} S(x_1, x_2) &= \int_0^\infty \exp\big\{-\big(\Lambda_1(x_1) + \Lambda_2(x_2) + z(x_1 + x_2)\big)\big\}\, g(z)\, dz \\ &= H_1(x_1)\, H_2(x_2)\, L_Z(x_1 + x_2), \end{aligned}$$

where $H_1(x_1) = e^{-\Lambda_1(x_1)}$, $H_2(x_2) = e^{-\Lambda_2(x_2)}$ and $L_Z(.)$ is the Laplace transform of Z.

Thus, the conditional density of Z given $T_1 > x_1, T_2 > x_2$ is given by

$$g\big(z|\, T_1 > x_1, T_2 > x_2\big) \;=\; \frac{\exp\big\{-z(x_1 + x_2)\big\}}{L_Z(x_1 + x_2)}\; g(z).$$

It can be verified that

$$E\big[Z \,|\, T_1 > x_1, T_2 > x_2\big] = \frac{-L_Z^{'}(x_1 + x_2)}{L_Z(x_1 + x_2)}$$

and

$$E\big[Z^2 | \, T_1 > x_1, T_2 > x_2\big] = \frac{L_Z^{''}(x_1 + x_2)}{L_Z(x_1 + x_2)}.$$

Hence

$$Var\big[Z \,|\, T_1 > x_1, T_2 > x_2\big] = \frac{L_Z^{''}(x_1 + x_2)}{L_Z(x_1 + x_2)} - \left(\frac{L_Z^{'}(x_1 + x_2)}{L_Z(x_1 + x_2)}\right)^2.$$

The above expressions yield

$$\theta(x_1, x_2) \;=\; 1 + \frac{\frac{L_Z^{''}(x_1+x_2)}{L_Z(x_1+x_2)} - \left(\frac{L_Z^{'}(x_1+x_2)}{L_Z(x_1+x_2)}\right)^2}{\left[\mu_1(x_1) - \frac{L_Z^{'}(x_1+x_2)}{L_Z(x_1+x_2)}\right]\left[\mu_2(x_2) - \frac{L_Z^{'}(x_1+x_2)}{L_Z(x_1+x_2)}\right]}. \tag{3.14}$$

We now present some examples

Example 3.4. Suppose Z has a gamma distribution with *pdf* given by (3.3). Also its Laplace transform and its derivatives are given in Example 3.1.

It can be verified that

$$\theta(x_1, x_2) \;=\; 1 + \frac{\alpha\beta^2}{\big[1+\beta(x_1+x_2)\big]^2}\left[A(x_1, x_2) + \left\{\frac{\alpha\beta}{\big[1+\beta(x_1+x_2)\big]}\right\}^2\right]^{-1}, \tag{3.15}$$

where

$$A(x_1, x_2) \;=\; \mu_1(x_1)\, \mu_2(x_2) + \frac{\alpha\beta}{\big[1+\beta(x_1+x_2)\big]}\, \big(\mu_1(x_1) + \mu_2(x_2)\big).$$

Thus $\theta(x_1, x_2) > 1$. Also as $x_1 \to \infty$ or $x_2 \to \infty$, $\theta(x_1, x_2) \to 1$. It is symmetric in x_1 and x_2 and is a decreasing function of x_1 or x_2.

Remark 3.3. Note that, in the multiplicative case, the value of $\theta(x_1, x_2)$ is independent of x_1 and x_2; see Hanagal (2011, page 83).

Example 3.5. Suppose Z has inverse Gaussian distribution with *pdf* given by (3.5). Also its Laplace transform and its derivatives are given in Example 3.2

It can be verified that

$$\theta(x_1, x_2) \;=\; 1 + \frac{a\big[b^2 + 2a\,(x_1 + x_2)\big]^{-3/2}}{A(x_1, x_2) + \big[b^2 + 2a\,(x_1 + x_2)\big]^{-1/2}}, \tag{3.16}$$

where

$$A(x_1, x_2) \;=\; \mu_1(x_1)\,\mu_2(x_2) + \big[b^2 + 2a\,(x_1 + x_2)\big]^{-1/2}\big(\mu_1(x_1) + \mu_2(x_2)\big).$$

Thus $\theta(x_1, x_2) > 1$. Also as $x_1 \to \infty$ or $x_2 \to \infty$, $\theta(x_1, x_2) \to 1$. It is symmetric in x_1 and x_2 and is a decreasing function of x_1 or x_2.

Example 3.6. Suppose Z has positive stable distribution with *pdf* given by (3.7). Also its Laplace transform and its derivatives are given in Example 3.3

It can be verified that

$$\theta(x_1, x_2) \;=\; 1 + \frac{\alpha(1-\alpha)\,(x_1 + x_2)^{\alpha-2}}{A(x_1 + x_2) + \alpha^2(x_1 + x_2)^{2\alpha-2}}, \tag{3.17}$$

where

$$A(x_1, x_2) \;=\; \mu_1(x_1)\,\mu_2(x_2) + \alpha(x_1 + x_2)^{\alpha-1}\big(\mu_1(x_1) + \mu_2(x_2)\big).$$

Thus $\theta(x_1, t_2) > 1$. Also as $x_1 \to \infty$ or $x_2 \to \infty$, $\theta(x_1, x_2) \to 1$. It is symmetric in x_1 and x_2 and is a decreasing function of x_1 or x_2.

Bivariate Gamma Correlated Additive Hazards Rate Model

Suppose Y_0, Y_1 and Y_2 are independent random variables and $Z_1 = Y_0 + Y_1$, $Z_2 = Y_0 + Y_2$. Then Z_1 and Z_2 are correlated.

The conditional survival function is given by

$$S\big(x_1, x_2 | Z_1 = z_1, Z_2 = z_2\big) \;=\; H_1(x_1)\,H_2(x_2)\,e^{-(z_1x_1 + z_2x_2)}.$$

We follow the notations and assumptions given in section 2.1. Here Z_1 and Z_2 have been taken with the same marginal distribution, but correlated.

This means that $Var(Z_1) = Var(Z_2) = \sigma_Z^2$ (say). Also the correlation coefficient between Z_1 and Z_2 will be denoted by ρ_Z.

We have

$$\rho_{Z_1,Z_2}(x_1,x_2) = \frac{Var_{Y_0}(x_1,x_2)}{\sigma_{Z_1}(x_1,x_2)\,\sigma_{Z_2}(x_1,x_2)}.$$

Under our assumptions

$$\alpha_0 = \rho_Z/\sigma_Z^2, \quad \alpha = (1-\rho_Z))/\sigma_Z^2, \quad \beta = 1/\sigma_Z^2,$$

$$Var_{Y_0}(x_1,x_2) = \frac{\alpha_0\sigma_Z^2}{\big[1+\sigma_Z^2(x_1+x_2)\big]^2} = \frac{\rho_Z\,\sigma_Z^2}{\big[1+\sigma_Z^2(x_1+x_2)\big]^2},$$

$$Var_{Y_i}(x_1,x_2) = \frac{\alpha\sigma_Z^2}{\big[1+\sigma_Z^2 x_i)\big]^2}, \quad i=1,2.$$

These give

$$\begin{aligned} Var_{Z_i}(x_1,x_2) &= Var_{Y_o}(x_1,x_2) + Var_{Y_i}(x_1,x_2) \\ &= \frac{\rho_Z\sigma_Z^2\big[1+\sigma_Z^2(x_i)\big]^2 + (1-\rho_Z)\,\sigma_Z^2\big[1+\sigma_Z^2(x_1+x_2)\big]^2}{\big[1+\sigma_Z^2(x_i)\big]^2\big[1+\sigma_Z^2(x_1+x_2)\big]^2}, \quad i=1,2. \end{aligned}$$

Thus

$$\begin{aligned} \rho_{Z_1,Z_2}(x_1,x_2) &= \frac{Var_{Y_0}(x_1,x_2)}{\sigma_{Z_1}(x_1,x_2)\,\sigma_{Z_2}(x_1,x_2)} \\ &= \frac{\rho_Z\big[\big(1+\sigma_Z^2(x_1)\big)\big(1+\sigma_Z^2(x_2)\big)\big]}{\Big[\prod_{i=1}^{i=2}\big\{\rho_Z[1+\sigma_Z^2(x_i)]^2 + (1-\rho_Z)[1+\sigma_Z^2(x_1+x_2)]^2\big\}\Big]^{1/2}}. \end{aligned} \tag{3.18}$$

Now

$$\begin{aligned} &E\big(Z_i|\,T_1>x_1, T_2>x_2\big) = \\ &\quad = E\big(Y_0|\,T_1>x_1,T_2>x_2\big) + E\big(Y_i|\,T_1>x_1,T_2>x_2\big) \\ &\quad = \frac{\rho_Z\big[1+\sigma_Z^2(x_i)\big] + (1-\rho_Z)\big[1+\sigma_Z^2(x_1+x_2)\big]}{\big[1+\sigma_Z^2(x_i)\big]\big[1+\sigma_Z^2(x_1+x_2)\big]}, \quad i=1,2. \end{aligned} \tag{3.19}$$

Using the above expressions, the $\mathrm{CV}_{Z_i}(x_1,x_2)$ is given by

$$\big[CV_{Z_i}(x_1,x_2)\big]^2 = \frac{\rho_Z\sigma_Z^2\big[1+\sigma_Z^2(x_i)\big]^2 + (1-\rho_Z)\,\sigma_Z^2\big[1+\sigma_Z^2(x_1+x_2)\big]^2}{\Big\{\rho_Z\big[1+\sigma_Z^2(x_i)\big] + (1-\rho_Z)\big[1+\sigma_Z^2(x_1+x_2)\big]\Big\}^2}, \quad i=1,2. \tag{3.20}$$

Using the expressions of $\rho_{Z_1,Z_2}(x_1,x_2)$, $CV_{Z_1}(x_1,x_2)$ and $CV_{Z_2}(x_1,x_2)$, $\theta(x_1,x_2)$ can be obtained.

Shared Frailty Model

In this case $\rho_Z = 1$ and hence $\rho_{Z_1,Z_2}(x_1, x_2) = 1$ and the expression for $\theta(x_1, x_2)$ simplifies to

$$\theta(x_1, x_2) = 1 + \sigma_Z^2.$$

4. SOME APPLICATIONS

In medical and epidemiological studies, the primary object is to study the effect of concomitant information on the time to event such as death or recurrence of a disease. Cox proportional hazard model is commonly used in the analysis of survival time data.

As has been indicated earlier, there is some amount of unobserved heterogeneity among individuals that is not accounted for by the Cox model. Failing to account this form of heterogeneity between individuals may lead to distorted results. Models, which account for this form of unobserved heterogeneity, are known as frailty models. The models are formulated based on the idea that individuals who are most frail will experience the event of interest earlier than others.

Price and Manatunga (2000) analyzed the leukemia patients data. In this data, leukemia patients receive either an allogenic transplant or an autologous transplant. Patients are followed and time to recurrence is recorded. They applied, cure models, frailty models and frailty mixture models to analyze this data. Specifically, the cure models, gamma frailty, gamma frailty mixture, inverse Gaussian frailty, inverse Gaussian mixture and compound Poisson models are utilized to model the data.

Xue and Ding (1999) applied the bivariate frailty model to inpatients mental health data. One frailty is used to represent heterogeneity across all hospital stays and another to represent heterogeneity across all community stays. These two frailties are jointly distributed. They show that this model offers much more flexibility than the univariate frailty model in modelling heterogeneity for the analysis of bivariate survival times.

Hens *et al.* (2009) considered multisera data on hepatitis A and B. They applied the bivariate correlated gamma frailty model for type I interval censored data. They showed that applying a shared rather than a correlated frailty model to this cross-sectionally collected serological data on hepatitis A and B leads to biased estimate for the baseline hazards and variance parameters. Weinke *et al.* (2003) point out that the shared frailty explains correlation within clusters. However, it does have some limitations.

Wienke *et al.* (2003) applied the correlated gamma frailty model to fit bivariate time to event (occurrence of breast cancer) data. They fitted the model for left truncated and right truncated censored data and the analysis accounts for heterogeneity as well as insusceptible (cure fraction) in the study population. This approach includes the shared gamma frailty model as a special case. The correlated gamma model provides a specific parameter for correlation between the two frailties. They also observed that individual frailties in twin pairs could not be observed, but their correlation could be estimated by application of the gamma frailty model.

Weinke *et al.* (2006) used three correlated frailty models to analyze survival data by assuming gamma, log-normal and compound Poisson distributed frailty. All approaches allow to deal with right censored data and account for heterogeneity as well as non susceptible (cure fraction) in the study population. Breast cancer incidence data of Swedish twin pairs illustrate the practical relevance of the models, which are used to estimate the cure fraction and the correlation between the frailties of the twin partners.

We have described some applications of frailty models and correlated frailty models. For more applications, the reader is referred to the bibliography in these papers and the books on frailty models.

5. SOME CONCLUSION AND COMMENTS

Multivariate survival distributions are used in the analysis of life spans of related individuals. An important class of such distributions can be derived by using the concept of random hazards. The randomness is modeled as a frailty random variable having an appropriate distribution. This paper presents a general bivariate correlated frailty model and unifies various results available in the literature. A bivariate gamma correlated frailty model is studied. Clayton's association measure is derived for the general model under study. Proportional hazards as well as additive hazards bivariate frailty model is investigated along with several examples. We hope that the results presented here will be found useful for researchers dealing with various problems involving frailty.

ACKNOWLEDGMENTS

The author is thankful to the referees for some useful comments which enhanced the presentation.

REFERENCES

[1] Agresti, A.; Caffo, B. and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency and possible remedies, *Computational Statistics and Data Analysis*, **47**, 639–653.

[2] Anderson, J.E.; Louis, T.A. and Holm, N. (1992). The dependent association measures for bivariate survival distributions, *Journal of the American Statistical Association,* **87**, 641–650.

[3] Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of family tendency in chronic disease incidence, *Biometrika*, **65**, 141–151.

[4] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of Royal Statistical Society, series B*, **34**(13), 187–220.

[5] Duchateau, L. and Janssen, P. (2008). *The Frailty Model*, Springer Verlag, NY.

[6] Eriksson, F. and Scheike, T. (2015). Additive Gamma frailty models with applications to competing risks related individuals, *Biometrics*, **71**, 677–686.

[7] Gorfine, M. and Hsu, L. (2011). Frailty based competing risks model for multivariate survival data, *Biometrics*, **67**, 415–426.

[8] Gupta, R.C. and Gupta, R.D. (2009). General frailty model and stochastic orderings, *Journal of Statistical Planning and Inference*, **139**, 3277–3287.

[9] Gupta, R.C. and Gupta, R.D. (2010). Random effect survival models and stochastic comparisons, *Journal of Applied Probability*, **47**, 426–440.

[10] Gupta, R.C. and Kirmani, S.N.U.A. (2006). Stochastic comparisons in frailty models, *Journal of Statistical Planning and Inference*, **136**, 3647–3658.

[11] Hanagal, D.D. (2011). *Modeling Survival Data Using Frailty Models*, Chapman and Hall, Boca Raton, FL.

[12] Hanagal, D.D. and Bhambure, S.M. (2016). Modeling bivariate survival data using shared inverse Gaussian frailty model, *Communications in Statistics — Theory and Methods*, **45**(17), 4969–4987.

[13] Hanagal, D.D. and Dabade, A.D. (2015). Comparison of shared frailty models for kidney infection data under exponential power baseline distribution, *Communications in Statistics — Theory and Methods*, **44**, 5091–5108.

[14] Heckman, J.J. and Singer, B. (1984). The identifiability of the proportional hazard model, *Rev. Econom. Stud. Li*, 231–241.

[15] Hens, N.; Wienke, A.; Aerts, M. and Molenberghs, G. (2009). The correlated and shared gamma frailty model for bivariate current status data: an illustration for cross-sectional serological data, *Statistics in Medicine*, **28**, 2785–2800.

[16] Hougaard, P. (1984). Lifetable methods for heterogeneous populations: distributions describing the heterogeneity, *Biometrika*, **71**, 75–83.

[17] Hougaard, P. (1991). Modeling heterogeneity in survival data, *Journal of Applied Probability*, **28**, 695–701.

[18] Hougaard, P. (1995). Frailty models for survival data, *Lifetime Data Analysis*, **1**, 255–273.

[19] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer, New York.

[20] Korsgaard, I.R. and Anderson, A.H. (1998). The additive genetic gamma frailty model, *Scandinavian Journal of Statistics*, **25**, 255–269.

[21] Marshall, A.W. and Olkin, I. (1988). Families of multivariate distributions, *Journal of the American Statistical Association*, **83**, 834–841.

[22] Oakes, D. (1989). Bivariate survival models induced by frailties, *Journal of the American Statistical Association*, **84**, 497–493.

[23] Price, D.L. and Manatunga, A.K. (2000). Modelling survival data with a cure fraction using frailty models, *Statistics in Medicine*, **20**, 1515–1527.

[24] Silva, G.L. and Amaral Turkman, M.A. (2004). Bayesian analysis of an additive survival model with frailty, *Communications in Statistics — Theory and Methods*, **33**, 2517–2533.

[25] Tomazella, V.L.; Louzada-Neto, F. and Silva, G.L. (2006). Bayesian modeling of recurrent events data with an additive gamma frailty distribution and a homogeneous Poisson process, *Journal of Statistical Theory and Applications*, **5**(4), 417–429.

[26] Vaupel, J.W.; Manton, K.G. and Sttalard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, **16**(3), 439–454.

[27] Wienke, A.; Lichtenstein, P. and Yashin, I.I. (2003). Bivariate frailty model with cure fraction for modelling correlations in diseases, *Biometrics*, **59**, 1178–1183.

[28] Wienke, A.; Licantelli, I. and Yashin, A.I. (2006). The modeling of a cure fraction in bivariate time to event data, *Austrian Journal of Statistics*, **35**(1), 67–76.

[29] Wienke, A. (2010). *Frailty Models In Survival Analysis*, Chapman & Hall/CRC, Boca Raton.

[30] Xue, X. and Ding, A.Y. (1999). Assessing heterogeneity and correlation of failure times with bivariate frailty model, *Statistics in Medicine*, **18**, 907–918.

[31] Yashin, A.I. and Iachine, I.A. (1995a). Survival of related individuals: an extension of some fundamental results of heterogeneity analysis, *Mathematical Population Studies*, **5**(4), 321–339.

[32] Yashin, A.I. and Iachine, I.A. (1995b). How long can humans live? Lower bound for biological limit of human longevity calculated from Danish twin data using correlated frailty model, *Mechanisms of Ageing and Development*, **80**, 147–169.

[33] Yashin, A.I.; Vaupel, J.W. and Iachine, I.A. (1995). Correlated individual frailty: an advantageous approach to survival analysis of bivariate data, *Mathematical Population Studies*, **5**(2), 145–159.

[34] Yin, G. and Ibrahim, J.G. (2005). A class of Bayesian shared gamma frailty models with multivariate failure time data, *Biometrics*, **61**, 208–216.

REVSTAT – STATISTICAL JOURNAL

Background

Statistics Portugal (INE, I.P.), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, publishing three times a year papers containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided to publish papers also in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work.

At the time, the Editorial Board was mainly composed by Portuguese university professors, being now composed by national and international university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal.

In 2001, the *Revista de Estatística* published three volumes special issue containing extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

The name of the Journal has been changed to REVSTAT - STATISTICAL JOURNAL, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

— The editorial policy will focus on publishing research articles at the highest level in the domains of Probability and Statistics with emphasis on the originality and importance of the research.

— All research articles will be refereed by at least two persons, one from the Editorial Board and another external.

— The only working language allowed will be English. — Four volumes are scheduled for publication, one in January, one in April, one in July and the other in October.

Aims and Scope

The aim of REVSTAT is to publish articles of high scientific content, in English, developing innovative statistical scientific methods and introducing original research, grounded in substantive problems.

REVSTAT covers all branches of Probability and Statistics. Surveys of important areas of research in the field are also welcome.

Abstract and Indexing Services

The REVSTAT is covered by the following abstracting/indexing services:

- Current Index to Statistics
- Google Scholar
- Mathematical Reviews
- Science Citation Index Expanded
- Zentralblatt für Mathematic

Instructions to Authors, special-issue editors and publishers

The articles should be written in English and may be submitted in two different ways:

- By sending the paper in PDF format to the Executive Editor (revstat@ine.pt) and to one of the two Editors or Associate Editors, whose opinion the author wants to be taken into account, together to the following e-mail address: revstat@fc.ul.pt

- By sending the paper in PDF format to the Executive Editor (revstat@ine.pt), together with the corresponding PDF or PostScript file to the following e-mail address: revstat@fc.ul.pt.

Submission of a paper means that it contains original work that has not been nor is about to be published elsewhere in any form.

Manuscripts (text, tables and figures) should be typed only in black on one side, in double-spacing, with a left margin of at least 3 cm and with less than 30 pages. The first page should include the name, institution and address of the author(s) and a summary of less than one hundred words, followed by a maximum of six key words and the AMS 2000 subject classification.

Authors are obliged to write the final version of accepted papers using LaTeX, in the REVSTAT style. This style (REVSTAT.sty), and examples file (REVSTAT.tex), which may be download to PC Windows System (Zip format), Macintosh, Linux and Solaris Systems (StuffIt format), and Mackintosh System (BinHex Format), are available in the REVSTAT link of the Statistics Portugal Website: http://www.ine.pt/revstat/inicio.html

Additional information for the authors may be obtained in the above link.

Accepted papers

Authors of accepted papers are requested to provide the LaTeX files and also a postscript (PS) or an acrobat (PDF) file of the paper to the Secretary of REVSTAT: revstat@ine.pt.

Such e-mail message should include the author(s)'s name, mentioning that it has been accepted by REVSTAT.

The authors should also mention if encapsulated postscript figure files were included, and submit electronics figures separately in .tiff, .gif, .eps or .ps format. Figures must be a minimum of 300 dpi.

Editorial Board