ISSN 1927-7032 (Print)
ISSN 1927-7040 (Online)

International Journal of Statistics and Probability

Vol. 2, No. 1 February 2013

CANADIAN CENTER OF SCIENCE AND EDUCATION

Editorial Board

Contents

International Journal of Statistics and Probability; Vol. 2, No. 1; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

A Competitor of the Kolmogorov–Smirnov Test for the Goodness-of-fit Problem

Claudio G. Borroni[1] & Paola M. Chiodini[1]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy

Correspondence: C. G. Borroni, Department of Statistics and Quantitative Methods, University of Milano-Bicocca, via Bicocca degli Arcimboldi 8, Milano 20126, Italy. Tel: 39-2-6448-3120. E-mail: claudio.borroni@unimib.it

Received: September 19, 2012 Accepted: November 25, 2012 Online Published: December 19, 2012

doi:10.5539/ijsp.v2n1p1 URL: http://dx.doi.org/10.5539/ijsp.v2n1p1

Abstract

The classical goodness-of-fit problem, in the case of a null continuous and completely specified distribution, is faced by a new version of the Girone–Cifarelli test (see Girone, 1964; Cifarelli, 1974 & 1975). This latter test was introduced for the two-sample problem and showed a substantial gain of power over other common tests based on the empirical distribution function, notably over the Kolmogorov–Smirnov test. First, the problem of the re-definition of the Girone–Cifarelli test-statistic is considered, by reviewing the literature on the subject. A classical remark by Anderson (1962) is shown to be useful to choose the integrating function in the newly defined test-statistic. The sample properties of such a test-statistic are then studied. A table of critical values is obtained by simulation; moreover, the asymptotic null distribution is considered and its accuracy as an approximation of the finite distribution is discussed. Finally, a simulation study, considering a wide set of distributions mostly used in applications, is conducted to compare the proposed test with its classical competitors. The study gives some indications to locate such situations where the Girone-Cifarelli test performs at its best, notably over the Kolmogorov–Smirnov test.

Keywords: goodness-of-fit tests, empirical distribution function, Girone–Cifarelli test, nonparametric statistical methods

1. Introduction

A random sample $x_1, \ldots, x_n$ is drawn from a population X with continuous distribution function F, to test the null hypothesis $H_0 : F(x) = F_0(x)$ against the alternative $H_1 : F(x) \neq F_0(x)$, $x \in \Re$, where F_0 is completely specified. This common goodness-of-fit problem is usually faced by three classes of tests: the chi-square test, the tests based on spacings and the tests based on the empirical distribution function (edf). In this latter class several test-statistics can be considered, usually by adapting their versions for the two-sample problem.

The most known test based on the edf F_n is surely the Kolmogorov–Smirnov test, which rejects H_0 for large values of the test-statistic

$$K_n = \sup_{t \in (-\infty, +\infty)} |F_0(t) - F_n(t)|. \tag{1}$$

As known, other test-statistics can be defined by considering the square of the difference $|F_0(t) - F_n(t)|$, like in the Cramér–Von Mises test

$$C_n = n \int_{-\infty}^{+\infty} [F_0(t) - F_n(t)]^2 \, \mathrm{d}F_0(t). \tag{2}$$

Notice that in the above considered test-statistics F_0, a continuous model, is compared with F_n, which has discontinuities at $x_1, \ldots, x_n$. However, in (1) the supremum of the difference $|F_0(t) - F_n(t)|$ is taken, while in (2) the squared difference $[F_0(t) - F_n(t)]^2$ is integrated with respect to the continuous function F_0. Because of these latter choices, no particular care is needed in the definition of the value taken by the edf at its points of discontinuity. This means that one can use the usual definition

$$F_n(x) = \frac{i}{n}, \qquad \text{for } x_{(i)} \leq x < x_{(i+1)} \quad (i = 0, \ldots, n), \tag{3}$$

(where $x_{(1)}, \ldots, x_{(n)}$ denotes the ordered sample, $x_{(0)} = -\infty$ and $x_{(n+1)} = +\infty$), which makes F_n to be right-continuous, or equivalently set

$$F_n(x_{(i)}) = \frac{i-c}{n} \qquad (i = 1,\dots,n), \tag{4}$$

(where c is chosen in [0,1]), so that F_n can take every value of its jump at $x_{(i)}$ $(i = 1,\dots,n)$.

Turning back to (2), the function with which the squared difference $[F_0(t)-F_n(t)]^2$ is integrated could be substituted by the edf itself. This choice allows to simplify the test-statistic as

$$C_n^{'} = n\int_{-\infty}^{+\infty} [F_0(t) - F_n(t)]^2 \,\mathrm{d}F_n(t) = \sum_{i=1}^{n} [F_0(x_{(i)}) - F_n(x_{(i)})]^2, \tag{5}$$

but the definition of $F_n(x_{(i)})$ becomes now relevant. However, Anderson (1962) pointed out that, when $c = 1/2$ is taken in (4), the test-statistics C_n and $C_n^{'}$ are equivalent, as the former can be also written as

$$C_n = \sum_{i=1}^{n} \left[F_0(x_{(i)}) - \frac{i-1/2}{n}\right]^2 + \frac{1}{12n}. \tag{6}$$

Besides such a latter equivalence, setting $c = 1/2$ in (4) is, as a matter of fact, a natural choice. Indeed, forcing the edf to take the mid-point of its jump at $x_{(i)}$ seems less arbitrary than choosing any other value in the jump (including the extremes i/n and $(i-1)/n$, $i = 1,\dots,n$.

Notice again that any choice of $F_n(x_{(i)})$, made to give a final form to $C_n^{'}$ in (5), does not affect the usual definition of the edf in the open intervals $(x_{(i)}, x_{(i+1)})$, $i = 1,\dots,n-1$. However, in the literature some modifications of the edf in such intervals were also proposed. For instance, Green and Hegazy (1976) pointed out that when the edf is re-defined as

$$F_n^{'}(x) = \frac{i+1/2}{n+1} \qquad \text{for } x_{(i)} < x < x_{(i+1)} \qquad (i = 1,\dots,n-1), \tag{7}$$

the criterion C_n in (2) reduces, up to a multiplicative constant, to

$$\sum_{i=1}^{n} \left[F_0(x_{(i)}) - \frac{i}{n+1}\right]^2, \tag{8}$$

which is shown to lead to a powerful test under some circumstances. Notice that the test-statistic in (8) can be also obtained from $C_n^{'}$ in (5) by re-defining accordingly the value of the edf at its discontinuities, that is by setting

$$F_n^{'}(x_{(i)}) = \frac{i}{n+1} \qquad (i = 1,\dots,n), \tag{9}$$

which is again the mid-point of the jump of $F_n^{'}$ at $x_{(i)}$. Other modifications of the definition of the edf in the open intervals $(x_{(i)}, x_{(i+1)})$ are known. By noticing that the term $i/(n+1)$ is actually the expectation of $F_0(x_{(i)})$ under the null hypothesis, Pyke (1959) proposed a new version of the Kolmogorov–Smirnov criterion (1), which in turn induces a further modification of the definition of the edf (see also Brunk, 1962).

The above remarks will be used in this paper to propose a goodness-of-fit version of the Girone–Cifarelli test, which was mainly studied for the two-sample problem. The definition of the test-statistic for goodness-of-fit purposes raises some questions which will be addressed in the next section, where the sample properties of the newly proposed test-statistic will be also analyzed. Section 3 will report some results of a simulation study, where the proposed test is compared with its most important competitors based on the edf. Section 4 will provide a real-data example and some conclusions.

2. Definition of the Test-statistic

Girone (1964) proposed a test for the equality of two populations X and Y, based on the statistic

$$(m+n)\int_{-\infty}^{+\infty} |F_n(t) - G_m(t)| \,\mathrm{d}H_{m+n}(t), \tag{10}$$

where F_n, G_m and H_{m+n} denote respectively the edf's of a n-sample from X, a m-sample from Y and the pooled $(m+n)$-sample. The test was actually originally proposed in the special case $n = m$ and its sample properties were studied by Cifarelli (1974 & 1975). Generalizations for the case $n \leq m$ were proposed by Goria (1972), by Borroni (2001) and independently by Schmid and Trede (1995). For the two-sample problem, the Girone–Cifarelli test proved to be superior to other common tests, notably the Kolmogorov–Smirnov test, under a wide set of circumstances. This fact is far from being unexpected, as in (10) the whole behavior of the difference $|F_n(t) - G_m(t)|$ is considered, while in the Kolmogorov–Smirnov test just its supremum is taken.

A goodness-of-fit version of the Girone–Cifarelli test would result useful. Using the same settings as in section 1, the edf F_n of the single n-sample is now to be compared with the null model F_0. The function with respect to which the difference $|F_n(t) - F_0(t)|$ is to be integrated could then be the null model F_0 or the edf F_n. As above remarked, this latter choice highly simplifies the structure of the test statistic, as

$$n \int_{-\infty}^{+\infty} |F_0(t) - F_n(t)| \, \mathrm{d}F_n(t) = \sum_{i=1}^{n} |F_0(x_{(i)}) - F_n(x_{(i)})|; \tag{11}$$

as a consequence, the definition of the value taken by the edf at its discontinuities becomes relevant. Following the above suggestion by Anderson (1962) for $C_n^{'}$, we can then take

$$F_n(x_{(i)}) = \frac{i - 1/2}{n} \quad (i = 1, \ldots, n), \tag{12}$$

and define

$$A_n^{'} = \sum_{i=1}^{n} \left| F_0(x_{(i)}) - \frac{i - 1/2}{n} \right|. \tag{13}$$

The sample properties of $A_n^{'}$ are easily derived from its two-sample equivalent. First of all notice that, being F_0 a continuous model, the variables $F_0(x_{(i)})$, $i = 1, ..., n$, are uniform over [0,1] and hence $A_n^{'}$ is distribution-free under H_0. For small sample sizes, the null distribution of $A_n^{'}$ can then be determined by simulation, as pointed out in the next section. Moreover, following Cifarelli (1975), $n^{(-1/2)} A_n^{'}$ is asymptotically distributed as the r.v.

$$\int_0^1 |w(\tau)| \, \mathrm{d}\tau \Big| w(1) = 0, \tag{14}$$

where $\{w(\tau), t \in [0,1]\}$ denotes the Brownian motion in [0,1]. A tabulation of the quantiles of (14) is found in Johnson and Killeen (1983); see also Shepp (1982 & 1991) and Takács (1993).

Differently from $C_n^{'}$, $A_n^{'}$ is not equivalent to the statistic obtained by using F_0 as an integrating function in (11). This is shown by considering that

$$A_n = \int_{-\infty}^{+\infty} |F_0(t) - F_n(t)| \, \mathrm{d}F_0(t) = \frac{1}{2} \sum_{i=1}^{n} \left[\left| F_0(x_{(i)}) - \frac{i-1}{n} \right| \left(F_0(x_{(i)}) - \frac{i-1}{n} \right) - \left| F_0(x_{(i)}) - \frac{i}{n} \right| \left(F_0(x_{(i)}) - \frac{i}{n} \right) \right]. \tag{15}$$

Schmid and Trede (1996) considered $\sqrt{n} A_n$ as a test-statistic and reported a small simulation study to evaluate its performance. They concluded that the power of A_n is quite close to the one of the Cramér–Von Mises test, without reporting situations where A_n performs definitely nor uniformly better than C_n. It should be pointed out that $A_n^{'}$, which has a rather simpler form, is not equivalent to A_n, even if the two tests have often similar powers. Consequently, the next section will first present some results of a simulation study without distinguishing between $A_n^{'}$ and A_n. In the following, some insights about the situations where the two tests are likely to perform differently will then be given. In a sense, the reported simulation study can be considered as an extension of the one by Schmid and Trade (1996), because it will be able to locate some alternatives where the test based on $A_n^{'}$, along with the one based on A_n, performs definitely better than the Cramér–Von Mises test.

3. Simulation Study

The first task to develop a goodness-of-fit test based on $A_n^{'}$ is to determine its critical values. As above mentioned, being F_0 completely specified and continuous, the transformation $F_0(X)$ gives a Uniform distribution over (0,1). Hence the null distribution of $A_n^{'}$ can be simulated by randomly generating a large number of samples from such a distribution, with a fixed size n. The critical values of the test can then be determined by computing the value taken by $A_n^{'}$ for each simulated sample as long as the related frequency distribution. For a selected range of sample sizes and some common significance levels, Table 1 reports the critical values of $n^{(-1/2)} A_n^{'}$ based on 10^6 simulated samples.

Table 1. Simulated critical values of $n^{(-1/2)}A_n^{'}$

α	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 50$	∞
0.01	0.7142	0.7364	0.7436	0.7465	0.7478	0.7518
0.05	0.5670	0.5747	0.5783	0.5791	0.5807	0.5821
0.10	0.4893	0.4942	0.4966	0.4972	0.4982	0.4993
0.15	0.4398	0.4439	0.4459	0.4462	0.4470	0.4480
0.20	0.4029	0.4064	0.4081	0.4088	0.4092	0.4103

As a term of comparison, the last column of Table 1 reports the critical values of the asymptotic distribution of $n^{(-1/2)}A_n^{'}$ (see section 2). The fast convergence to the asymptotic approximation can be easily appreciated. In order to get further indications about the accuracy of the asymptotic distribution and the sample sizes needed to use it, the simulated cdf's obtained for fixed values of n were compared with the asymptotic cdf, whose expression is found in Johnson and Killeen (1983). Figure 1 reports the results obtained for $n = 10$. As seen, the asymptotic cdf is very close to the "real" one, even if a certain difference is observed, especially for small values of the variable. However, one can claim that, to develop a goodness-of-fit test based on $A_n^{'}$, just the right tail of its null distribution is relevant. In effect, when the last part of the distribution is considered (say for such x so that $\Pr\{A_n^{'} \leq x\} > 0.8$) the finite cdf is rather close to the asymptotic cdf. To get into further details, Table 2 reports, for some selected sample sizes, the greatest absolute difference of the two cdf's and the same difference referred to the right tail of the distribution. From such a table, a minimum value of $n = 50$ is to be advised to get a correct approximation of the null distribution.

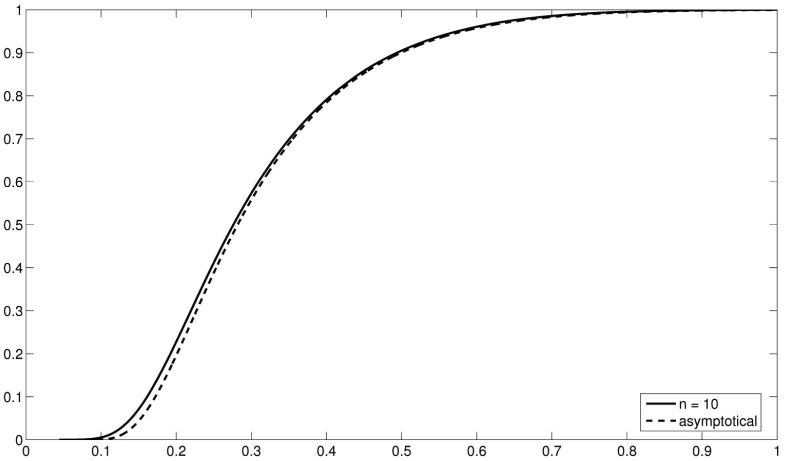


Figure 1. Comparison of the "real" cdf and the asymptotic cdf of $A_n^{'}$ under H_0 ($n = 10$)

Table 2. Greatest absolute difference between the "real" cdf and the asymptotic cdf of $A_n^{'}$ under H_0 (whole distribution and right tail)

n	whole distrib.	right tail
5	0.0626	0.0112
10	0.0330	0.0065
20	0.0173	0.0035
30	0.0123	0.0023
50	0.0068	0.0018
100	0.0042	0.0005

After computing the critical values of the test based on $A_n^{'}$, its power can be estimated by simulation as well. This section reports some results of a wide simulation study conducted at this aim. Notice that the power of a goodness-of-fit test will depend on the model F_0 chosen under H_0 as long as on the real cdf of X under H_1, which will be denoted as F_1. Generally F_1 will belong to a family of distributions containing F_0 itself, which is hence obtained by an appropriate choice of the parameter(s) of the family. In this paper we will focus on three models for H_0, mostly used in applications: the standard Normal distribution, the unit exponential distribution and the uniform distribution on the unit interval. For each null model, F_1 will belong to three different families of distributions containing F_0.

Consider first the standard Normal as a null model. A suitable family for F_1 could be the skew-normal (SN) distribution (see Azzalini, 1985) with density:

$$f(x) = 2\,\phi(x)\,\Phi(ax), \qquad x \in \Re, \tag{16}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and the cdf of a standard Normal respectively. The parameter $a \in \Re$ regulates the skewness of the distribution, thus giving a standard Normal if set to zero. To this end, using family (16) for F_1 in a simulation study, can result in an useful analysis of such situations where the researcher needs to test normality against possible asymmetries of data. It is known, however, that data may depart from normality due to heavy-tailedness. To simulate such latter situations, the Student's T density can be used as a family for F_1 :

$$f(x) = \frac{\Gamma\left(\frac{a+1}{2}\right)}{\sqrt{a\pi}\,\Gamma\left(\frac{a}{2}\right)}\left(1 + \frac{x^2}{a}\right)^{-\frac{a+1}{2}}, \quad x \in \Re, \tag{17}$$

($\Gamma(\cdot)$ denotes the gamma function). The family (17) gives only symmetric distributions with heavy tails, such phenomenon being reduced by increasing the parameter $a > 0$; as known, the family converges to the normal distribution when $a \to \infty$. Finally, to simulate such cases where the normality of data depends on the application of the central limit theorem, one can choose for F_1 the gamma (GA) density with unit scale:

$$f(x) = \frac{1}{\Gamma(a)}\,x^{a-1}\mathrm{e}^{-x}, \quad x > 0. \tag{18}$$

As an effect of the above theorem, when $a \to \infty$, family (18) gives a normal density (which can be then standardized to be consistent with the null model F_0). However, in applications, a may not be large enough to guarantee a good convergence; the researcher may then need a powerful test to detect such a failed convergence.

Table 3. Simulated powers when the null model is a standard Normal distribution ($\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	K_n	C_n	D_n
10	null	.0101	.0099	.0099	.0103
		.0500	.0506	.0502	.0506
		.1007	.1010	.1013	.1002
10	SN(1)	**.1989**	.1575	.1925	.1707
		.4503	.3849	.4392	.4088
		.5948	.5308	.5841	.5555
10	SN(1.5)	**.3320**	.2743	.3291	.2755
		.6460	.5739	.6399	.5908
		.7850	.7243	.7807	.7431
10	T(1.5)	.0312	.0243	.0284	**.4673**
		.1117	.0933	.1044	**.6088**
		.1948	.1772	.1843	**.6868**
10	T(1.25)	.0369	.0302	.0339	**.5883**
		.1272	.1062	.1197	**.7080**
		.2185	.2007	.2076	**.7703**
10	GA(2)	.0119	**.0203**	.0158	.0146
		.0687	**.0841**	.0748	.0768
		.1363	.1485	.1415	**.1525**

Table 3 (continued). Simulated powers when the null model is a standard Normal distribution ($\alpha = 0.01, 0.05, 0.1$).

n	H_1	$A_n^{'}$	K_n	C_n	D_n
10	GA(1.5)	.0132	**.0240**	.0179	.0169
		.0769	**.0946**	.0839	.0879
		.1508	.1653	.1556	**.1716**
25	SN(0.5)	**.2022**	.1491	.1904	.1904
		.4215	.3494	.4067	.4054
		.5501	.4809	.5385	.5350
25	SN(1)	**.6637**	.5441	.6471	.6302
		.8716	.7989	.8627	.8585
		.9317	.8894	.9272	.9251
25	T(1.75)	.0335	.0274	.0313	**.6016**
		.1299	.1107	.1219	**.7532**
		.2321	.1986	.2160	**.8246**
25	T(1.5)	.0392	.0347	.0374	**.7334**
		.1528	.1329	.1434	**.8468**
		.2717	.2355	.2543	**.8957**
25	GA(2)	.0296	**.0455**	.0363	.0316
		.1230	**.1403**	.1314	.1395
		.2171	.2295	.2201	**.2520**
25	GA(1.5)	.0398	**.0592**	.0483	.0427
		.1585	.1745	.1754	**.1817**
		.2703	.2692	.2698	**.3192**
100	SN(0.25)	**.2427**	.1775	.2304	.2400
		.4684	.3886	.4523	.4626
		.5904	.5198	.5777	.5859
100	SN(0.5)	**.8476**	.7331	.8315	.8439
		.9514	.9040	.9453	.9507
		.9762	.9505	.9730	.9760
100	T(2)	.0728	.0677	.0675	**.9398**
		.2909	.2545	.2708	**.9839**
		.4779	.4245	.4563	**.9931**
100	T(1.75)	.1005	.1005	.0953	**.9811**
		.3682	.3421	.3225	**.9959**
		.5697	.5291	.5558	**.9984**
100	GA(2)	.1959	.2275	.2179	**.2678**
		.4719	.4442	.4711	**.6476**
		.6408	.5777	.6278	**.8400**
100	GA(1.5)	.3151	.3245	.3351	**.4569**
		.6350	.5839	.6327	**.8505**
		.7863	.8208	.7805	**.9635**

Table 3 reports the results of a set of simulations, each based on 10^5 replications, for the null standard normal model. Some selected alternative distributions, all belonging to the above described families, are chosen. Table 3 reports the powers of the tests based on $A_n^{'}$, K_n and C_n. Another classical goodness-of-fit test is also considered:

the Anderson–Darling test,

$$D_n = \int_{-\infty}^{+\infty} [F_n(t) - F_0(t)]^2 \, \frac{1}{F_0(t)[1 - F_0(t)]} \, \mathrm{d}F_0(t); \tag{19}$$

here the squared difference $[F_n(t) - F_0(t)]^2$ is weighted to get more sensibility in the tails of the distributions. The powers reported in Table 3 were obtained by fixing three different values of the significance level α: 0.01, 0.05 and 0.1 (for each entry, the corresponding powers are listed in the latter order; the best power is highlighted too). It seems, however, that the performance of none of the considered tests is really affected by the choice of α. Moreover, the first row of Table 3 reports the estimated actual significance level, which is always very close to the nominal one, even for a small sample size as $n = 10$ (similar results, not reported here for the sake of brevity, were obtained for larger sample sizes). Notice that, for each considered alternative distribution, the values of the related parameter were set to allow a relevant comparison of the estimated powers; this need implies, incidentally, that the same value of the parameter cannot be chosen for all sample sizes in most cases. However, Table 3 (along with the following tables) was built so that at least one same value of the parameter is chosen for adjacent sample sizes. Table 3 emphasizes that, when used as a test of normality, the Girone–Cifarelli test has a good power against some kinds of alternatives. More specifically, the test outperforms all the other considered tests (and notably the Kolmogorov–Smirnov and the Cramér–Von Mises test) when the alternative distribution belongs to the skew-normal class. The superiority of the Girone–Cifarelli test for skewed alternatives seems indeed to be a general rule, at least among the considered tests, as further evidenced in the following simulations. When normality is to be tested against heavy tailedness, like for the Student's T alternatives considered in Table 3, the performance of the Girone–Cifarelli test gets worse, notably over the Anderson–Darling test. This result is far from being unexpected, but it has to be underlined that $A_n^{'}$ still keeps its superiority over the Cramér–Von Mises test (and the Kolmogorov–Smirnov test). The superiority of the Anderson–Darling test still characterizes Gamma alternatives. In this chance, however, the Girone–Cifarelli test gets worse even over the Cramér–Von Mises test and the Kolmogorov–Smirnov test. A global evaluation of Table 3 shows that, as expected, the performances of the considered tests become similar when the sample size increases, even if all the above conclusions still hold. Notably, $A_n^{'}$ outperforms the other considered tests for skew-normal alternatives, as D_n does for Student's T and Gamma alternatives. However, in this latter case, the Girone–Cifarelli test seems to grow better over its competitors as the sample size increases.

Another set of simulations was conducted by setting the unit exponential distribution as a null model. This assumption is typical for many datasets in reliability analyses. In this kind of applications, exponentiality is often to be tested against some more complicate distributional assumptions. To this end, a natural choice for F_1 is again the gamma (GA) density with unit scale. When $a = 1$, (18) reduces to the unit exponential. Another family of distributions, mostly used in reliability analyses as well, is the Weibull (W) density with unit scale:

$$f(x) = a\, x^{a-1}\, \mathrm{e}^{-x^a}, \qquad x > 0, \tag{20}$$

which was used as a family for F_1, after noticing that it reduces to the unit exponential when $a = 1$. Finally, a third family was used to shape the alternative hypothesis:

$$f(x) = (1 + a\, x)^{-\left(1+\frac{1}{k}\right)}, \qquad x > 0, \tag{21}$$

that is the generalized-Pareto (GP) density with zero location and unit shape. (21) gives the unit exponential density as $a \to 0$. The best results for the Girone–Cifarelli test were obtained for the Gamma alternatives, as shown by Table 4, which has the same settings as Table 3. $A_n^{'}$ outperforms all the other considered tests, notably the Cramér–Von Mises test. The Anderson–Darling test has generally a worse power than $A_n^{'}$, even if it becomes its main competitor as n increases. It has to be emphasized that the simulated powers reported for Gamma alternatives in Table 4 cannot be compared with the ones reported in Table 3, as the null distribution is quite different in the two sets of simulations. More specifically, when the null model is the Normal distribution, the power of each considered test is a decreasing function of the parameter a in (18); conversely, when the null model is the unit exponential distribution, the power is an increasing function, at least if $a > 1$. This fact explains why, even if the sample size and the value of a may coincide, Table 4 and Table 3 report quite different values of the estimated powers. Turning to other distributions considered in Table 4, one can notice that the above conclusions are reversed for alternatives of the generalized-Pareto kind, as D_n outperforms here all the other tests; $A_n^{'}$ has a similar power to the one of the Cramér–Von Mises test, but it seems to worsen as n increases. The Weibull alternatives evidence a problem of bias for some tests under some circumstances; apart from this fact, this case resembles the Student's T alternatives of Table 3: except for $n = 5$, the test based on D_n has definitely the best power, but $A_n^{'}$ clearly outperforms the Cramér–Von Mises and the Kolmogorov–Smirnov tests.

Table 4. Simulated powers when the null model is a unit exponential distribution ($\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	K_n	C_n	D_n
10	null	.0094	.0100	.0097	.0095
		.0494	.0495	.0494	.0492
		.1001	.0992	.1001	.0989
10	GA(1.5)	**.1515**	.1152	.1437	.1444
		.3560	.2891	.3428	.3420
		.4827	.4229	.4710	.4684
10	GA(2)	**.6191**	.4863	.5928	.6048
		.8388	.7511	.8222	.8269
		.9088	.8507	.8988	.9004
10	W(2)	.0021	**.0144**	.0049	.0010
		.0618	**.0984**	.0712	.0360
		.1951	**.2036**	.1903	.1315
10	W(3)	.0038	**.0501**	.0105	.0007
		.2587	**.2889**	.2579	.1469
		.5906	.4925	.5440	.4554
10	GP(0.35)	.0269	.0215	.0257	**.0871**
		.0946	.0829	.0908	**.2048**
		.1659	.1486	.1585	**.3019**
10	GP(0.45)	**.0411**	.0323	.0337	.1469
		.1179	.1018	.1119	**.2830**
		.1892	.1964	.1830	**.3821**
25	GA(1.25)	**.1098**	.0818	.1018	.1039
		.2772	.2249	.2645	.2649
		.3882	.3334	.3764	.3765
25	GA(1.5)	**.4973**	.3752	.4711	.4811
		.7339	.6351	.7149	.7226
		.8281	.7538	.8140	.8201
25	W(1.75)	.0224	**.0554**	.0326	.0235
		.2296	.2205	.2199	**.2317**
		.4582	.3806	.4291	**.4618**
25	W(2)	.0708	**.1160**	.0818	.0689
		.4462	.3758	.4174	**.4549**
		.7093	.5743	.6680	**.7231**
25	GP(0.35)	.0422	.0344	.0395	**.1456**
		.1339	.1153	.1271	**.3028**
		.2168	.2007	.2152	**.4181**
25	GP(0.45)	.0647	.0531	.0616	**.2599**
		.1743	.1564	.1696	**.4433**
		.2705	.2531	.2672	**.5603**

Table 4 (continued). Simulated powers when the null model is a unit exponential distribution ($\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	K_n	C_n	D_n
100	GA(1.15)	**.1843**	.1332	.1737	.1829
		.3945	.3228	.3812	.3916
		.5202	.4504	.5064	.5144
100	GA(1.25)	**.5705**	.4317	.5476	.5675
		.7873	.6902	.7700	.7868
		.8705	.7996	.8578	.8696
100	W(1.5)	.2994	.2756	.2920	**.4657**
		.7534	.6165	.7201	**.8719**
		.9046	.7945	.8842	**.9605**
100	W(1.75)	.8391	.7054	.8199	**.9463**
		.9880	.9462	.9832	**.9986**
		.9982	.9850	.9977	**.9999**
100	GP(0.35)	.1441	.1424	.1504	**.5048**
		.3378	.3432	.3488	**.7313**
		.4754	.4844	.4865	**.8288**
100	GP(0.45)	.2359	.2712	.2604	**.7616**
		.4807	.5287	.5124	**.9028**
		.6251	.6670	.6529	**.9461**

In the simulations reported in Table 4, both the null and the alternative distributions are skewed. To consider other cases where a null symmetric model is to be tested against skewed alternatives, like for the above standard Normal case, a third set of simulations is finally reported. The uniform distribution on the unit interval $(0, 1)$ is used as a null model. Some "modifications" of the uniform density are considered as alternatives. The first has density

$$f(x) = \begin{cases} a\ (2x)^{a-1} & 0 \leq x \leq 0.5, \\ a\ (2-2x)^{a-1} & 0.5 \leq x \leq 1, \end{cases} \tag{22}$$

(where $a > 0$) and it is labeled as MU; the second,

$$f(x) = (1-2a)^{-1}, \qquad a < x < 1-a, \tag{23}$$

is essentially a "compressed" uniform (CU) distribution over the interval $(a, 1-a)$, where $0 \leq a \leq 1/2$. Both densities reduces to the uniform distribution on the unit interval when $a = 0$. They were drawn from the study conducted by Schmid and Trede (1996). To complete such an investigation, a third alternative family is here considered for F_1:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} x^{a-1}(1-x)^{b-1}, \qquad 0 < x < 1. \tag{24}$$

The Beta (B) density in (24) reduces to the uniform distribution on $(0, 1)$ when $a = b = 1$. In the reported simulation study, b is then set to 1 and $a > 0$ is left to vary. Notice that, as a grows over 1, the distribution becomes more skewed, thus giving exactly the needed kinds of alternatives.

Table 5. Simulated powers when the null model is a uniform distribution on the unit interval ($\alpha = 0.01, 0.05, 0.1$)

n	H_1	A_n'	K_n	C_n	D_n
10	null	.0100	.0097	.0101	.0098
		.0507	.0501	.0501	.0503
		.1000	.1001	.0998	.1004
10	MU(4.5)	.0001	.0047	.0001	.0000
		.2705	**.2974**	.2263	.1042
		.6944	.4963	.6357	.5301
10	MU(5)	.0001	.0051	.0001	.0000
		.3862	.2588	.3115	.1569
		.7963	.5793	.7432	.6469
10	CU(0.3)	.0001	.0032	.0003	.0000
		.1621	.0890	.1219	.0480
		.6449	.3176	.5732	.4564
10	CU(0.35)	.0002	.0039	.0005	.0000
		.5598	.2214	.4939	.2308
		.9814	.8019	.9931	.9715
10	B(2)	**.1951**	.1556	.1898	.1686
		.4501	.3830	.4378	.4085
		.5941	.5289	.5829	.5548
10	B(3)	**.6371**	.5259	.6251	.5755
		.8780	.8048	.8689	.8451
		.9408	.8998	.9372	.9243
25	MU(3.5)	**.3870**	.2725	.3411	.3740
		.9392	.7778	.9174	**.9474**
		.9922	.9349	.9873	**.9940**
25	MU(4)	**.6321**	.4205	.5783	.6197
		.9874	.8985	.9812	.9906
		.9992	.9811	.9982	**.9994**
25	CU(0.2)	**.0242**	.0234	.0202	.0232
		.4149	.2214	.3596	**.5457**
		.7733	.5318	.7540	**.9305**
25	CU(0.225)	**.0574**	.0432	.0466	.0620
		.6697	.4173	.6478	.8581
		.9374	.8782	.9554	.9985
25	B(1.5)	**.1833**	.1386	.1740	.1661
		.4043	.3390	.3935	.3831
		.5395	.4766	.5280	.5210
25	B(2)	**.6650**	.5449	.6490	.6319
		.8694	.7972	.8621	.8561
		.9314	.8894	.9265	.9244
100	MU(1.5)	.0648	.0901	.0668	**.1444**
		.4222	.3506	.3963	**.5674**
		.6748	.5539	.6434	**.7775**

Table 5 (continued). Simulated powers when the null model is a uniform distribution on the unit interval ($\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	K_n	C_n	D_n
100	MU(1.75)	.3958	.3005	.3642	**.6086**
		.8715	.7177	.8446	**.9441**
		.9659	.8822	.9549	**.9871**
100	CU(0.1)	.0805	.0474	.0635	**.3127**
		.4913	.2932	.4456	**.9256**
		.7670	.5883	.7563	**.9996**
100	CU(0.12)	.2246	.1208	.1959	**.7489**
		.7880	.6817	.8031	**.9996**
		.9500	.9990	.9712	**.9999**
100	B(1.25)	**.2458**	.1851	.2351	.2413
		.4830	.4078	.4704	.4801
		.6123	.5431	.6008	.6099
100	B(1.5)	.8400	.7374	.8276	**.8410**
		.9523	.9100	.9474	**.9544**
		.9777	.9563	.9755	**.9796**

Table 5 reports some results of this last set of simulations. The alternatives of kind (22) and (23) evidence that all the considered tests suffer from a problem of bias, to which the Anderson–Darling test seems to be the most exposed. A second remark is that the performance of $A_n^{'}$ is similar to the one of C_n, even if the former has almost everywhere a higher estimated power. Both the Girone–Cifarelli and the Cramér–Von Mises test are outperformed by the Anderson–Darling test for as large sample sizes as $n = 100$ (for alternatives (23) even from $n = 25$). These conclusions add few extra details to the ones obtained by Schmid and Trede (1996) for the test based on A_n in (15), which has actually a performance similar to the Girone–Cifarelli test. However, the alternatives of the Beta class represent a considerable addition in the evaluation of such tests of uniformity: the power of $A_n^{'}$, as long as the one of A_n (unreported), is here steadily over the one of the other considered tests, notably over the one of C_n, with some minor exceptions for D_n. Even if Table 5 reports just the results for selected values of the parameter a in (24), the simulation study showed the Girone–Cifarelli test to be uniformly more powerful than the Cramér–Von Mises test for $a > 1$.

The discussion of Table 5 raises an important issue to be considered before giving some general conclusions in the next section. As stated from the very beginning, the Girone–Cifarelli test performs often similarly to the test based on A_n in (15); this fact resulted clearly from the conducted simulation study and it is essentially the reason why no separate results about A_n are reported in the above discussion. However, the two test-statistics $A_n^{'}$ and A_n are not equivalent, as evidenced by the following simple decomposition:

$$A_n = \frac{2}{n} \sum_{i \in \overline{\mathcal{A}}} \left| F_0(x_{(i)}) - \frac{i-1/2}{n} \right| + 2 \sum_{i \in \mathcal{A}} \left[\left(F_0(x_{(i)}) - \frac{i}{n} \right) \left(F_0(x_{(i)}) - \frac{i-1}{n} \right) + \frac{1}{2n^2} \right], \tag{25}$$

where $\mathcal{A} \equiv \left\{ i : \frac{i-1}{n} < F_0(x_{(i)}) < \frac{i}{n} \right\}$. Notice that the set $\mathcal{A}$ is not empty (and thus $A_n^{'}$ and A_n are not equivalent) as long as the empirical distribution function is not dominated by the null model F_0 (or conversely). Hence the possible differences in the powers of $A_n^{'}$ and A_n are likely to be observed when the alternative distribution does not dominate the null model (or conversely), a fact that can be partially guaranteed by letting the two distributions have the same location. A last set of simulations was then conducted where the alternative distribution was forced to have the same mean of the null model. In effect, some of the above-reported alternative distributions do not guarantee such a requirement. In addition, small values of the sample size were chosen, as the effect of the second summand in (25) is likely to decrease with n. Table 6 reports some results when $A_n^{'}$ and A_n are used to test unit exponentiality against other skewed alternatives, a situation which proved to be good for both tests against their classical competitors. On the average, $A_n^{'}$ turns out to perform still similarly to A_n, even if there are cases where the difference in their powers becomes relevant. Notice that, with some minor exceptions, the Girone–Cifarelli test has never a lower power with respect to A_n. $A_n^{'}$ outperforms A_n for Gamma and notably for generalized-Pareto alternatives. The Weibull case is less definite, as $A_n^{'}$ has only a minor advantage over A_n and not for very small same sizes.

Table 6. Comparison between the powers of $A_n^{'}$ and A_n when the null and the alternative distributions have the same location (H_0 = unit exponential, $\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	A_n
5	GA(2)	.0625	.0615
		.1534	.1504
		.2464	.2411
5	GA(3)	.0885	.0874
		.1909	.1866
		.3120	.3043
5	GA(4)	.1058	.1045
		.2167	.2106
		.3601	.3503
5	W(3)	.0000	.0000
		.0354	.0382
		.2207	.2278
5	W(4)	.0000	.0000
		.0353	.0459
		.3403	.3526
5	W(5)	.0000	.0000
		.0401	.0572
		.4746	.4870
5	GP(0.45)	.2889	.2866
		.4406	.4323
		.5734	.5635
5	GP(0.47)	.3162	.3138
		.4659	.4574
		.6032	.5937
5	GP(0.49)	.3428	.3409
		.4953	.4864
		.6322	.6226
7	GA(2)	.0624	.0618
		.1665	.1645
		.2778	.2749
7	GA(3)	.0849	.0839
		.2275	.2242
		.3745	.3709
7	GA(4)	.1030	.1019
		.2745	.2698
		.4475	.4418
7	W(3)	.0014	.0015
		.1873	.1876
		.5175	.5172

Table 6 (continued). Comparison between the powers of $A_n^{'}$ and A_n when the null and the alternative distributions have the same location (H_0 = unit exponential, $\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	A_n
7	W(4)	.0007	.0007
		.3404	.3422
		.7563	.7557
7	W(5)	.0005	.0004
		.5029	.5108
		.8934	.8941
7	GP(0.45)	.3406	.3386
		.5568	.5522
		.6969	.6937
7	GP(0.47)	.3689	.3666
		.5898	.5853
		.7280	.7246
7	GP(0.49)	.3999	.3971
		.6216	.6167
		.7578	.7550
9	GA(2)	.0643	.0638
		.1834	.1815
		.3042	.3014
9	GA(3)	.0933	.0927
		.2655	.2623
		.4319	.4276
9	GA(4)	.1147	.1140
		.3358	.3314
		.5240	.5189
9	W(3)	.0163	.0168
		.4135	.4128
		.7501	.7486
9	W(4)	.0297	.0305
		.6895	.6878
		.9370	.9365
9	W(5)	.0546	.0539
		.8668	.8649
		.9874	.9871
9	GP(0.45)	.4183	.4166
		.6565	.6532
		.7814	.7778
9	GP(0.47)	.4593	.4579
		.6923	.6891
		.8124	.8097
9	GP(0.49)	.4943	.4927
		.7261	.7232
		.8381	.8355

4. A Real-Data Example and Some Conclusions

Before drawing some conclusions, a simple example where the test proposed in this paper is applied to real data is reported. The "warp breaks" dataset in Pearson (1963) has fast become a term of comparison of the results of various goodness-of-fit tests where the null distribution F_0 is completely specified. In this case, one has to test if the places where some warp breaks occur on a loom can be considered as uniformly distributed on the whole length of the warp. More specifically, the following distances of $n = 20$ breaks from the beginning of the warp are recorded: 30, 36, 104, 286, 291, 658, 893, 955, 1149, 1195, 1208, 1240, 1277, 1282, 1363, 1384, 1421, 1477, 1504, 1510 (see Pearson, 1963 for further details) and the ratios of these distances with respect to the total length (1520) are considered; a goodness-of-fit test is then applied to verify if such a sample of ratios comes from a population with unit uniform distribution. The observed value of $A_n^{'}$ is 2.9964, so that $n^{(-1/2)}A_n^{'} = 0.6633$ and, by looking at Table 1, the null hypothesis is to be rejected at the 5%-level but not at the 1%-level. More specifically, by using the simulated null distribution of $A_n^{'}$, a p-value 0.0213 is obtained. As a term of comparison, the p-values of the other considered tests are: 0.0090 for the Kolmogorov–Smirnov test, 0.0156 for the Cramér–Von Mises test and 0.0110 for the Anderson–Darling test. The results of all tests are then consistent, even if some differences can be appreciated.

This paper presents a simulations study which gives some new insights about goodness-of-fit tests based on the empirical distribution function. The main conclusion is that a good analysis should never neglect tests based on the averaged absolute difference $|F_n(t) - F_0(t)|$. The tests based on $A_n^{'}$ and A_n will both serve at this aim, even if the former can give some slight advantages over the latter, at least for small sample sizes. Moreover, the test-statistic $A_n^{'}$ has a rather simple form and it can be computed very easily. A second important conclusion is that $A_n^{'}$ (and A_n) has very often a different performance from the one of C_n, which is based on the averaged squared difference $[F_n(t) - F_0(t)]^2$. The reported simulations give a good evidence of such alternatives where $A_n^{'}$ outperforms C_n. It seems, specifically, that this happens more frequently for skewed alternatives. Concerning the Kolmogorov–Smirnov test K_n, which takes into consideration the supremum and not an average of the difference $|F_n(t) - F_0(t)|$, the reported study shows that there are few practical situations where it performs better than the other considered tests, and notably than $A_n^{'}$. The superiority, under some circumstances, of the Girone–Cifarelli test over the Kolmogorov–Smirnov test has been evidenced, in effect, in other studies concerning the two-samples problem. As a last issue, one can claim that the real competitor of $A_n^{'}$ (and similarly for A_n) is the Anderson–Darling test D_n, rather than K_n or C_n. The discussion in this paper shows that there are cases where D_n outperforms all other considered tests and that it leaves $A_n^{'}$ as a second best. These are mainly cases of alternatives with heavy tails, probably thanks to the weighting function in the definition of D_n. An important element of a future research could then be to evaluate the effect of the introduction of suitable weighting functions in the definition of $A_n^{'}$ as well.

References

Anderson, T. W. (1962). On the distribution of the two-sample Cramér-von Mises criterion. *Ann. Math. Statist., 33*, 1148-1159. http://dx.doi.org/10.1214/aoms/1177704477

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist., 12*, 171-178.

Borroni, C. G. (2001). Some notes about nonparametric tests for the equality of two populations. *Test, 10*, 147-159. http://dx.doi.org/10.1007/BF02595829

Brunk, H. D. (1962). On the range of the difference between hypothetical distribution function and Pyke's modified empirical distribution function. *Ann. Math. Statist., 33*, 525-532. http://dx.doi.org/10.1214/aoms/1177704578

Cifarelli, D. M. (1974). Intorno alla distribuzione del test di G. Girone. *Giornale degli Economisti e Annali di Economia, 33*, 283-308.

Cifarelli, D. M. (1975). Contributi intorno ad un test per l'omogeneità tra due campioni. *Giornale degli Economisti e Annali di Economia, 34*, 233-249.

Girone, G. (1964). Su un indice di omogeneità di due distribuzioni del tipo dell'indice semplice di dissomiglianza. *Atti della XXIV Riunione Scientifica della SIS*, 53-78.

Goria, M. (1972). Some generalizations of Girone's test statistic. *Annali dell'Istituto di Statistica - Università di Bari*, 53-72.

Green, J. R., & Hegazy, Y. A. S. (1976). Powerful modified-EDF goodness-of-fit tests. *J. Amer. Statist. Assoc., 71*, 204-209. http://dx.doi.org/10.1080/01621459.1976.10481516

Johnson, B., & Killeen, T. (1983). An explicit formula for the c.d.f. of the L1 norm of the Brownian bridge. *Ann. Prob., 11*, 807-808. http://dx.doi.org/10.1214/aop/1176993528

Pearson, E. S. (1963). Comparison of tests for randomness of points on a line. *Biometrika, 50*, 315-325.

Pyke, R. (1959). The supremum and infimum of the Poisson process. *Ann. Math. Statist., 30*, 568-576. http://dx.doi.org/10.1214/aoms/1177706269

Schmid, F., & Trede, M. (1995). A distribution free test for the two sample problem for general alternatives. *Comput. Stat. Data Anal., 20*, 409-419. http://dx.doi.org/10.1016/0167-9473(95)92844-N

Schmid, F., & Trede, M. (1996). An L1-variant of the Cramér-von Mises test. *Stat. Probabil. Lett., 26*, 91-96. http://dx.doi.org/10.1016/0167-7152(95)00256-1

Shepp, L. A. (1982). On the integral of the absolute value of the pinned Wiener process. *Ann. Prob., 10*, 234-239. http://dx.doi.org/10.1214/aop/1176993926

Shepp, L. A. (1991). Acknowledgement of priority. *Ann. Prob., 19*, 1397. http://dx.doi.org/10.1214/aop/1176990351

Takács, L. (1993). On the distribution of the integral of the absolute value of the Brownian motion. *Ann. Appl. Probab., 3*, 186-197.

International Journal of Statistics and Probability; Vol. 2, No. 1; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Studies on the Probabilistic Model for Ship-Bridge Collisions

Shaowei Zhang[1]

[1] Highway College, Chang'an University, Shanxi Province, China

Correspondence: Shaowei Zhang, Highway College, Chang'an University, Zhuhong Road, Xi'an 710018, Shanxi Province, China. E-mail: zfjzsw@yahoo.com.cn

Received: October 19, 2012 Accepted: December 4, 2012 Online Published: December 19, 2012

doi:10.5539/ijsp.v2n1p16 URL: http://dx.doi.org/10.5539/ijsp.v2n1p16

Abstract

Shocking ship-bridge collisions indicate that there's large space in the previous bridge anti-collision technology research. There are several advantages in the risk-based anti-collision technology of the bridges. Thus the databases such as SpringerLink, Elsevier ScienceDirect and CNKI, the Chinese database, are included to collect literature for the purpose of examining the probabilistic models. Reviewing the current representative models, this paper argues some limitations in the models, such as the questionable applicability of models, the neglected affects of pier turbulent zones as well as some inaccuracies in the mathematical formulations. Accordingly, the paper revises the current models and also addresses increasing the representativeness of samples with sufficient experiments. This paper explores the topic for its potential applications, and aims to make some contribution to the references on the topic so as to popularize and promote the technology in a real sense.

Keywords: vessel-bridge collision, probabilistic model, limitations, further studies

1. Introduction

Shocking ship-bridge collisions indicate that bridge anti-collision technology research still has very large space in that the practicality and operability of the research results need further proof. Therefore, the author used PDF-Geni and Google as search engines to collect literature, with *bridge anti-collision* as key words. Meanwhile, the databases such as SpringerLink, Elsevier ScienceDirect and CNKI, the Chinese database, were also included. The literature review found that previous risk-based studies on the bridge anti-collision at home and abroad are comparatively deficient. In the case of optimizing the placement of the bridge sensors, the method proposed by Guo (2010) can not only alarm the collision between ships and bridges, but also can note down the data of accidents to evaluate the degree of damage in the bridge. However, studies like Guo's excessively focused on those mathematical models of the collision probability, underestimating the impact of the turbulent zone around piers on the collision probability. Accordingly, Jiang and Wang (2009) recommended using AASHTO model and LARSEN model if the calculation is adjusted to the domestic situation. Based on the previous researches at home and abroad, this paper reviews and comments on the present researches on the probabilistic models for ship-bridge collision, discusses the limitations of the studies, and addresses the corresponding improvements, attempting to make some contributions to the present literature.

2. Literature Review and Commentary

2.1 Literature Review

Bridge anti-collision technology is classified into passive technology and active technology and the domestic researches mostly focus on the passive one, for instance, setting the mechanical anti-collision device to reduce the impact of the collision between the ship and bridge. However, it is acknowledged that it's impossible to block all the collisions unless it depends on the bridge itself (Larsen, 1993; Vrouwenvelder, 1998). Furthermore, the bridge anti-collision device costs too much, for example, the cost of the flexible energy-absorbing anti-collision device for the main pier of *Zhanjiang Bay Bridge* in China remains twenty million RMB, which is an unacceptable cost for the regular bridges. The existing bridge anti-collision devices are restricted to the critical bridges in the dense waterway. In contrast, bridge anti-collision technology based on risk ideas has the advantage of preventing accidents in advance. Thus we should pay enough attention to the risk-based anti-collision research on the bridges, and studies on the probabilistic models for ship-bridge collision catch the author's eye.

The result of applying the AASHTO Method II design procedure is the calculation of an annual frequency of collapse for a given bridge. For critical bridges, the risk acceptance criterion is less than or equal to 0.0001, or once every ten-thousand years. For regular bridges, the acceptable risk is less than or equal to 0.001, or once every thousand years (AASHTO, 1994. *AASHTO LRFD Bridge Design Specification and Commentary*). Collision risk models consider the effects of the vessel traffic, the navigation conditions, the bridge geometry with respect to the waterway, and the bridge element strength with respect to the impact loads (Knott & Pruca, 2000). By reviewing the previous literature, 5 models are currently found to the most representative ones. These models taken from original works are listed as following, whose inaccuracies will be put off until the 3^{rd} section.

2.2.1 AASHTO Model

The 1991 AASHTO Specifications provide three methods (Methods I, II, and III) for designing a bridge while taking into account potential vessel impact. Method II is the only method presented in the 2001 AASHTO LRFD Bridge Design Specification, whose essential data include vessel description, speed and loading conditions, waterway geometry, navigable channel geometry, water depths etc. Under AASHTO Method II, bridges must be assigned an importance classification as a Regular or Critical bridge, based on society/survival demand and security/defense requirements (AASHTO: 2009. *Guide Specification and Commentary for Vessel Collision Design of Highway Bridges*). The equation for the calculation of an annual frequency of collapse for a given bridge is generally formulated as follows:

$$AF = N \cdot PA \cdot PG \cdot PC$$

where,

AF = The annual frequency of bridge element collapse due to vessel collision;

N = The annual number of vessels classified by type, size, and loading which can strike the bridge element;

PA = The probability of vessel aberrancy;

PG = The geometric probability of a collision between an aberrant vessel and a bridge pier or span;

PC = The probability of bridge collapse due to a collision with an aberrant vessel.

To provide an alternative means for calculating the probability of aberrancy, the 2001 AASHTO Specifications allow this probability to be approximated using the equation below:

$$PA = BR \cdot R_B \cdot R_C \cdot R_{XC} \cdot R_D$$

where,

PA = The probability of aberrancy;

BR = The aberrancy base rate;

R_B = The correction factor for bridge location;

R_C = The correction factor for current acting parallel to vessel transit path;

R_{XC} = The correction factor for crosscurrents acting perpendicular to vessel transit path;

R_D = The correction factor for vessel traffic density.

The AASHTO model uses dynamic analysis to determine the force of ships and also provides a simplified way to design a probability model to simulate the ship-bridge collision. The AASHTO model is based on the results of accidents, and the movement of ships is not related to the probability of vessel aberrancy (*PA*), or the geometric probability of a collision (*PG*). The probability calculation is larger than the truth value unless the probability of a collision not between the ship and the vessel is eliminated.

2.1.2 Larsen Model

In 1991, Larsen proposed the collision risk model at IABSE's annual conference (Larsen, 1993), which is expressed in the following form, where the first summation refers to all ship classes considered and the second summation refers to all bridge piers and superstructure spans:

$$F = \sum N_i \cdot P_{C,j} \cdot \sum P_{G,i,k} \cdot P_{F,i,k}$$

where,

F = Expected number of annual collisions to the bridge (bridge piers and/or superstructure);

N_i = Annual number of vessels belonging to a certain class (i) of the vessels passing the bridge;

$P_{c,j}$ = The "causation probability" related to the actual class of vessel (i);

$P_{G,i,k}$ = The "geometrical probability" or "rate of collision candidates" related to the actual class of vessel (i) and to the actual part (pier or span) of the bridge (k);

$P_{F,i,k}$ = The "failure probability" related to the actual class of vessels (i) and to the actual part of the bridge (k).

A probabilistic approach is based on a probabilistic model for the vessel impact force and a spatial stochastic model of the resistance properties of the bridge elements. Larsen model calculates the probability of bridge failure, which means not until $P_{F,i,k}$ being removed can the calculation truly represent the probability of the bridge collision (Jiang & Wang, 2009). As described in AASHTO model, the "causation probability" $P_{c,j}$ does not change with the sailing course. And the accidents are classified into the linear impact, meeting impact and random drifting impact, related to the angle of attack and the different failure modes of the bridge elements (e.g. crushing, rotation, sliding, etc.). Meanwhile, the linear impact can be subdivided into impacts on the axis of channel and those at the turns or bends in the navigation route. When applying the model at a certain river, we should get the gross impact probability by considering the ratio of the three situations in all accidents.

2.1.3 Eurocode Model

In 1997, Eurocode proposed a model to calculate the probability of the ship-bridge collision in volume 1. The model uses the centerline of the channel as X axis and parallel Y axis with the bridge axis, and the pier is located at $X = 0$ and $Y = d$ (Vrouwenvelder, 1998). Ship-bridge collision is considered as a non-homogeneous Poisson process, assuming that the error of Poisson process is $\lambda(x)$ so that the probability of the collision in a referenced period T can be expressed as follows:

$$P_C(T) = nTP_{na} \iint \lambda(x) P_C(x,y) f_s(y) dxdy$$

where,

$P_c(T)$ = The probability of not avoiding at least one collision within the reference period (usually 1 year);

n = The number of ships per time unit (traffic intensity);

T = The reference period (usually 1 year);

P_{na} = The probability that a collision is unavoided in spite of human intervention;

$\lambda(x)$ = The probability of a failure per unit traveling distance, determined with reference to data of previous accidents;

$P_c(x,y)$ = The probability of situations where a collision occurs with a given initial ship position (x, y);

$f_s(y)$ = The distribution of the ship position in the y-direction.

Eurocode and AASHTO Specifications share the similarity in the basic design philosophy. Eurocode 1, Part 2.7 refers in a note to ISO (Draft Proposal DP 10252): 1995. *Accidental Action due to Human Activities*, which specifies the representative value of an accidental action should be chosen in such a way that there is an assessed probability less than p=10^{-4} per year for one structure (Vrouwenvelder, 1998). Although the acceptable risk criterion is determined by each country government, but the acceptable annual frequency of collapse they recommend for the critical bridge is less than or equal to 1×10^{-4}, or once every ten-thousand years (Knott, 1998; AASHTO, 2009. *Guide Specification and Commentary for Vessel Collision Design of Highway Bridges*). Various collision risk models have been developed to achieve design acceptance criteria, while determining the risk acceptance criteria is based on the society's willingness to pay for the risk reduction.

2.1.4 KUNZI Model

Based on the variables describing the accidental course of the ship, a mathematical risk model was formulate by the German researcher Kunz (1998), in which a deviation on the maneuvering path with angle φ and the stopping distance x are chosen. Given the numerous affecting elements, the minimum distance x necessary for avoiding the pier should be a normal random variable. The collision model is outlined here in the following:

$$P(T) = N \cdot \int \frac{d\lambda}{ds} \cdot W_1(s) \cdot W_2(s) ds$$

where,

$P(T)$ = The probability of not avoiding at least one collision within the reference period (usually 1 year);

N = The number of ships per time unit (traffic intensity);

T = The reference period (usually 1 year);

$d\lambda/ds$ = The failure rate per travel unit;

$W_1(s)$ = The probability of collision course;

$W_2(s)$ = The probability not to come to a stop before collision to structure.

where,

$$W_1(s) = F_\varphi(\varphi_1) - F_\varphi(\varphi_2)$$

$$F_\varphi(\varphi) = \frac{1}{\sqrt{2\pi}\sigma_\varphi} \int_{-\infty}^{\varphi} \exp\left\{ \frac{(\varphi - \overline{\varphi})^2}{2\sigma_\varphi^2} \right\} d\varphi$$

where, $\bar{\varphi}, \sigma_\varphi$ are mean value and standard deviation of the angle φ between the planned course and the maneuvering course path;

$$W_2(s) = 1 - F_x(s)$$

$$F_x(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \int_{-\infty}^{x} \exp\left\{ \frac{(x - \overline{x})^2}{2\sigma_x^2} \right\} dx$$

where, $\bar{x}, \sigma_x$ are mean value and standard deviation of stopping distance x, referring to the distance between the ship and the pier when the ship detecting the danger of collision and taking urgent measures.

By calculating the probability $W_1(s)$ and $W_2(s)$ for each position along the approaching course of the ship, any probability of collision can be determined. The failure rate is mainly determined by accidents analysis, simulation, or by transferring such value from other technical systems (Galor, 2005). KUNZI model as well as Eurocode model focus on the process of the ship bridge collision. The former calculates the probability of a collision between the ship on a course to a bridge, while the latter does offer the mathematical equation for $P_C(x, y)$ in the given location (x, y). Therefore, the equation $W_1(s) \cdot W_2(s)$ in KUNZI model are recommended to use when calculating $P_c(x, y)$ in Eurocode model, meanwhile the distribution of the ship location in the y-direction $fs(y)$ should be taken into consideration when calculating the probability of collision. As a result, KUNZI model becomes a concrete formulation of the Eurodecode model. However, it is not so convenient to determine the probability of collision in Eurocode model and KUNZI model as to determine in AASUTO model and Larsen model (Jiang & Wang, 2009).

2.1.5 Dai Tongyu Simplified Model

Based on numerous experiments and data analyses, Dai et al. (2003) formulated a simplified model to calculate the probability of a collision, which applies more to the navigational conditions in China. It is hypothesized that the collision frequency F_i of ship class (i) is relevant to the probability of a collision p_i on the course with potential collisions and the value affecting collisions f_i, the model is then defined as follows,

$$F = \sum_i N_i \cdot f_i \cdot p_i$$

where p_i is determined based on the normal distribution of navigation courses. Based on the distribution of navigation courses of the ship passing the bridge, the mean value μ and the standard deviation σ are calculated in the following model:

$$p_i = \int_A^B \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

The probability of a collision to a bridge refers to the summation of the probabilities that passing ships come into collision with the pier and other structures of the bridge. The mathematical equation is formulated in the following:

$$F = \sum_i N_i \cdot f_i \cdot p_i = \sum_i N_i \cdot f_i \int_A^B \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

where,

N_i = The number of ship class (i) per time unit (traffic intensity);

f_i = The value affecting the collisions ship class (i), such as navigation course, current, weather, ship size, speed, direction etc.;

p_i = The probability of a collision on the course with potential collisions;

μ = The mean value of the location that a ship passes the axis of a bridge;

σ = The standard deviation of the location when a ship passing the axis of the bridge.

The feasibility and applicability of the simplified model has been proved by verifying the ship-bridge collision accidents of *Nanking Yangtse River Bridge* in China. Based on the relevant statistics of the waterway of the mentioned bridge, the value affecting the collisions f varies from 0.05 to 0.12. However, as far as the bigger ship sizes are concerned, the affecting value f may be a bit smaller.

2.2 Limitations in Relevant Models

AASHTO model and Larsen model attach attention to extreme situations, and the calculation focuses on the probability of the bridge destruction, while Eurocode model, Kunzi model and Dai Tongyun simplified model pay attention to all the accidents including the situations that the bridge is not destroyed. Excluding the failure probability, AASHTO Model and Larsen Model will be as practical as the other three models. However, these mathematical models have additional drawbacks in that their samples are not representative enough, the influence of the turbulent zones around piers is not considered in the calculation, and there're some mathematical inaccuracies in the equations.

2.2.1 Insufficient Samples and Doubtful Applicability of Models

Admittedly, excessive stress on the affecting factors is not significant because some of the factors do not affect a lot and even can be ignored. However, when the river system is different, the applicability of those models should be doubtful. Let's take Dai Tongyu simplified model as an example. In the case of *Huangshi Yangtse River Bridge* in Hubei with 20 ship-bridge collisions after it came into use, its hydrological conditions around piers is comparably more complex than those around *Nanking Yangtze River Bridge*. Whether Dai Tongyu simplified model can still be applied to this bridge or not obviously needs further consideration and verification. Dai Tongyu simplified model only verified its applicability in the middle and lower *Yangtze River* and was formulated only based on the hydrological conditions around *Nanking Yangtse River Bridge*. The upper reaches is fast-flowing, with a straight and smooth river way and a "V" font river valley, while the middle and lower reaches is mostly slow-flowing, with a winding river way and a "U" font river valley. Obviously, there is a significant difference between the hydrogeological conditions of the upper and lower reaches. In comparing the upper and lower reaches in just one river system, we do find that the net width of navigable channel, angle between the sailing direction and the axis of a bridge, the stopping distance have changed a lot. The applicability of the models is doubtful, let alone applying Dai Tongyu simplified models to a totally different river system such as *Great Canal* and *Yellow River*.

2.2.2 Neglected Impact of Turbulence Zones around Piers

When a current flows by the piers, there will be vortex which gives attraction to the surface layer around the piers. It's called the turbulence zones, whose width depends on the type of the pier as well as the size and shape of the river under the bridge. When the ship enters the turbulent zones, it will be exerted by an attraction which points to the pier. If we still use the present mathematical models to evaluate the risk regardless of the turbulence zones, we will underestimate the probability of the ship-bridge collision. Some domestic researchers simply include the width of the turbulence zones into the calculation (Gong, 2010), which may fall into the wrong idea that "any boat moving into any area of the turbulence zones will have a collision". Nowadays, the peripheral area of a turbulence zones perhaps can not make any difference to the ships with increasing weight and velocity, so counting the whole width of the turbulence zones without careful consideration can shorten the navigation span, which may cause problems to some bridges. To conclude, the relevant researches to date lack the accurate verification on the influence of turbulence zones on the calculation of collision probability.

2.2.3 Some Inaccuracies in the Mathematical Equations

There're some inaccuracies from the viewpoint of mathematics. Taking Dai Tongyu simplified model as an example, an index i should not be included in the formulation of p_i. The index i means different types of ships, but when calculating the value of p_i, the model uses data and courses of all the ships to get the value of μ and σ, which indicates that the value of p_i means no difference to different types of ships. Thus the model should be revised as follows.

$$F = \sum_i N_i \cdot f_i \cdot p = \sum_i N_i \cdot f_i \int_A^B \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Also, the sum in Larsen Model should be more clearly stated. The index i refers to different types of ships and k refers to different parts of a bridge, which should be indicated more directly. The model should be completed as follows.

$$F = \sum_i N_i \cdot P_{C,j} \cdot \sum_k P_{G,i,k} \cdot P_{F,i,k}$$

3. Ideas on Further Studies

AASHTO model is an empirical formula though most of its parameters are statistical. Those statistical parameters mainly focus on the main piers, which leads to the errors in calculating the collision possibility concerning the transition piers and the piers of approach bridges. KUNZI model and Eurocode model focus on the process of the accidents, without taking into account the wind speed, visibility and navigational aids so that the results of the calculation tend to be relatively larger. Accordingly, the paper makes the recommendations as follows.

3.1 Enlarging the Capacity of Samples

Vessel collision accidents to bridge structures are relatively rare and conditions differ from bridge to bridge. Therefore, the estimation of the risk of collision can not be based on vessel/bridge collisions alone. Collision risk models, stimulating potential collision scenarios are necessary (Larsen, 1993), thus simulating the collision accidents to enlarge the capability of samples is recommended hereby. The stimulation of collision consists of the computer-assisted technique as well as the realistic stimulation technique. The computer stimulation technique, such as FEM stimulation approach, can stimulate numerous characteristics such as the collision force, deformation of the structure or collision energy change. As for the realistic technique, we may choose one bridge which is going to be abandoned in each different river system and put them into a second use. Samples need to be representative, so that we can simulate the collisions with different vessel number (in the morning, at noon, in the afternoon) and in different situations (at night with light interference, upstream, downstream, different visibility etc.). To make the statistics more representative, we should relax the drivers or even deliberately distract the drivers. Getting too close to the piers or being fairly difficult to manipulate the ship when coming into the turbulence zones should be counted as most collisions have actually taken place on account of human errors. Of course, the experiment safety is to be guaranteed by some well planned protective measures.

3.2 Applying Revised Models in Bridge Design

Larsen (1993) and Vrouwenvelde (1998) addressed that risk assessment of the bridges should be based on the probabilistic models. Thereby the paper suggests using the newly revised models to calculate the bridge collision. As is pointed out, the impact of the turbulence zones should be included in the calculation of the probabilistic models. The parameter f_T is here used to represent the influence coefficient of the turbulence zones:

$$f_T = k \cdot f(D, \beta, v_1, v_2, h)$$

The parameter k represents the actual correction factor to influence the moving of the ships, D represents the size of the piers, β represents the angle between the moving direction of the river and the axis of the bridge, v_1 represents the velocity of the water flow in front of the piers, v_2 represents the velocity of the wind in front of the bridge and h represents the depth of the river around the piers. With the influence coefficient of the turbulence zones considered and removing the term of the failure probability, new models with better applicability are addressed in the following:

AASHTO model

$$AF = N \cdot PA \cdot PG \cdot f_T$$

Larsen model

$$F = \sum_i N_i \cdot P_{C,j} \cdot \sum_k P_{G,i,k} \cdot f_T$$

Eurocode model

$$P_C(T) = nTP_{na} \cdot f_T \cdot \iint \lambda(x) P_C(x, y) f_s(y) dx dy$$

KUNZI model

$$P(T) = N \cdot f_T \cdot \int \frac{d\lambda}{ds} \cdot W_1(s) \cdot W_2(s) ds$$

Dai Tongyu simplified model

$$F = \sum_i N_i \cdot f_i \cdot p \cdot f_T$$

From the design point of view, the bridge characteristics would be adjusted or the risk reduction requirements would be implemented until the risk acceptance is satisfied (Knott, 1998). The purpose of the risk assessment is to reduce the collision probability and provide theoretical support for the adjustment and perfecting of the bridge design. After the design proposal of a bridge is scheduled, the latest probability model should be used to simulate, analyze and predict all the possible bridge collisions so that the probability of a collision is minimized before putting the design proposal into construction. Likewise, anti-collision devices and better shipping management are also necessary after a bridge is constructed, for instance we can turn to alarming facilities.

4. Conclusive Remarks

Despite its good practicality, AASHTO model presents larger in the calculation results. Eurocode model features focusing on the process of the accidental action, in which a collision occurs when a vessel approaching the bridge becomes aberrant, or the aberrant vessel hits a bridge element, or the stricken bridge element fails. KUNZI model as well as Eurocode model merely focuses on the process of vessel bridge collision. Therefore, Jiang and Wang (2009) proposed to calculate the collision probability in AASHTO model or KUNZI model, on condition that some adjustments should be taken into consideration based on the domestic navigation conditions. On the basis of the previous researches, this paper has reached the following conclusions:

1) Analyzing the representative models, the paper has further discovered the questionable applicability of models, the neglected affects of pier turbulent zones in the models and some mathematical inaccuracies in the probabilistic models.

2) Accordingly, the paper has completed the probabilistic models with mathematical inaccuracies, and further revised the current models with the influence coefficient f_T aiming to improve the practicality of the probabilistic models.

3) This paper has also proposed increasing the representativeness of samples with sufficient experiments, the application of current researches into the design of bridges, and improving the system of shipping management with the aid of alarming facilities.

The paper has attempted to apply the more verified research findings to the anti-collision technology of the bridges so as to popularize and promote the technology in a real sense. Of course, the bridge anti-collision technology based on risk idea has its limitations. No matter how strong the risk idea-based anti-collision capacity is, even if a pier has the least probability to be impacted and the most accurate alarming systems, we do not have enough time to stop a collision when a ship is fairly close to that pier. Therefore, we still cannot delay the research on the anti-collision devices. Furthermore, ship owners have, in principle, the same interest as bridge owners, since the collision will bring damage and losses to both ship owners as well bridge owners (Manen & Frandsen, 1998). Thereby, only by improving the comprehensive anti-collision technology can we fundamentally ensure the safety of the bridge to fulfill their designed life, as well as the ship owner to escape the losses.

Acknowledgements

My sincere gratitude goes to Dr. Jiang Hua for his valuable advice on this paper, who received his bridge engineering PhD from Tongji University, and currently teaches in Chang'an University, Xi'an. My gratitude also goes to the anonymous reviewers of the journal for their inspiring and helpful advice.

References

Dai, T. Y., Liu, W. L., & Nie, W. (2003). Probability Analysis and Prediction of Ship Impacts against Bridges. *Journal of Harbin Engineering University, 1*, 23-29.

Deng, A. Y., Gao, J. D., & Du, Y. T. (2011). Risk Analysis of Vessel-Bridge Collision Based on AASHTO Model Algorithm. *World Bridges, 1*, 55-58.

Galor, W. (2005). The Movement of Ship in Water Areas Limited by Port Structures. *Annual of Navigation, 10*, 23-37.

Gong, T. (2010). Studies on the Probability of Ship-bridge Collision. Wuhan University of Technology.

Guo, Y. L. (2010). Monitoring-based Assessment of Bridges Subject to Ship Collision. The Hong Kong Polytechnic University.

Jiang, H., & Wang, J. J. (2009). Ascertaining Resistance Force of Bridge Components Against Vessel Collision Based on Risk Idea. *Structural Engineers, 6*, 67-71.

Knott, M. A. (1998). Vessel collision design codes and experience in the United States (pp. 75-84). In Gluver & Olsen (Eds.), *Ship Collision Analysis*. A. A. Balkema, Rotterdam. ISBN: 9054109629.

Knott, M., & Pruca, Z. (2000). Vessel Collison Design of Bridges (Section 60). In W. F. Chen, & L. Duanz (Eds.), *Bridge Engineering Handbook*. Boca Raton: CRC Press. ISBN: 0849374340.

Kunz, C. U. (1998). Ship bridge collision in river traffic analysis and design practice (pp. 13-22). In Gluver & Olsen (Eds.), *Ship Collision Analysis*. A. A. Balkema, Rotterdam. ISBN: 9054109629.

Larsen, O. D. (1993). Ship Collision with Bridges: Interaction between Vessel Traffic and Bridge Structures. *Structural Engineering Documents* (SED 4). Switzerland: IABSE. ISBN: 38574807933.

Manen, S. E., & Frandsen, A. G. (1998). Ship collision with bridges, review of accidents (pp. 3-11). In Gluver & Olsen (Eds.), *Ship Collision Analysis*. A. A. Balkema, Rotterdam. ISBN: 9054109629.

Tang, Y., Jin, Y. L., & Zhao, Z. Y. (2010). Comparison and Application of Probability Models of Ship-Bridge Collision. *Journal of Shanghai Ship and Shipping Research Institute, 1*, 28-33.

Vrouwenvelder, A. C. W. M. (1998). Design for ship impact according to Eurocode 1 Part 2.7 (pp. 123-132). In Gluver & Olsen (Eds.), *Ship Collision Analysis*. A. A. Balkema, Rotterdam. ISBN: 9054109629.

International Journal of Statistics and Probability; Vol. 2, No. 1; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Empirical Value at Risk for Weak Dependent Random Variables

Ali Kabui[1] & Samir Ben Hariz[1]

[1] Laboratoire de Statistique et Processus, Université du Maine, France

Correspondence: Ali Kabui, Laboratoire de Statistique et Processus, Université du Maine, 72 Av. Olivier Messiaen, 72085 Le Mans Cedex 9, France. E-mail: ali.kabui.etu@univ-lemans.fr

Received: October 7, 2012 Accepted: November 12, 2012 Online Published: December 19, 2012

doi:10.5539/ijsp.v2n1p24 URL: http://dx.doi.org/10.5539/ijsp.v2n1p24

Abstract

In this work, we study the empirical estimator of the Value at Risk (VaR for short) for weak dependent observations. Our approach uses the oscillation of the empirical process under hypothesis of moment's inequality. We provide general conditions which ensure the convergence of the empirical estimator of the VaR. We also prove a central limit theorem (CLT) for the difference. We perform some simulations for different sequences to illustrate our results. Finally, we apply the results for different sequences under assumptions of mixing or covariance.

Keywords: Value at Risk (VaR), modulus of continuity, empirical process, quantile function, moment's inequality, dependent random variables

1. Introduction

The Value at Risk VaR is a method to evaluate financial risks. It summarizes the risks of loss in a unique number and aggregating the risks of market through several classes of financial assets (stocks, bonds, etc.).

The VaR is a probabilistic measure of the possible loss for a given horizon. It represents a level of loss, for a financial position or a portfolio, which will be exceeded during a given period only with a chosen typically small probability.

The VaR is obviously neither the loss which one can expect nor the maximum loss which one may suffer, but a level of loss which will be exceeded only with a level of a fixed probability q.

Definition 1 (*P&L and loss function*) Let P_t be the value of a portfolio of assets at time t. Then the variation of the value of this portfolio over the interval $[t, t+T]$, is called the profit-and-loss (P&L) function:

$$\triangle P_t \equiv P_{t+T} - P_t,$$

and the function

$$X_t :\equiv -\triangle P_t$$

is called the loss function.

In practice, we decide to fix T (e.g. one day or one week), yet $\triangle P_t \equiv P_{t+1} - P_t$.

Definition 2 (*Value at Risk*) The Value at Risk $VaR\,(q)$ of a portfolio of assets for a period $[t, t+1]$ at the confidence level $q \in (0,1)$ is given by the smallest number x such that the probability that the loss X_t exceeds x is no larger than $(1-q)$. Formally

$$VaR(q) \equiv \inf\{x : \mathbb{P}(X_t > x) \leq 1-q\}$$

or

$$VaR\,(q) \equiv F_t^{-1}(q) = \inf\{x : F\,(x) \geq q\} := \xi. \tag{1.1}$$

where $F_t\,(x) = \mathbb{P}\,(X_t \leq x)\,, x \in \mathbb{R}$ is the distribution function of X_t and F_t^{-1} its quantile function.

Definition (1.1) clearly shows that the knowledge of the distribution function (in short df) of the r.v X can determine the $VaR\,(q)$. Often the function F is assumed to be normal. However a lot of financial practitioners use historical distributions which are far from being normally distributed (see e.g. Cont, 2001). Moreover, in general, the historical data have an intertemporally dependent structures. Indeed the assumption that the variables $(X_i)_{1\leq i\leq n}$ (denote the variations $(-\triangle P_i)_{1\leq i\leq n}$ in the value of a portfolio over the n periods) are i.i.d, is not easily satisfied in practice. Hence the feeling of the need of taking into account a possible dependence structure or an effect of memory in the observations. In order to model and measure this memory aspect in the data, we consider two cases: correlations or mixing coefficients.

So the main objective of this paper is to provide ways which allow to tackle the issue of estimation of the VaR in the cases where there is either a lack of parameterizations of F or some weak dependency among the data. To do so, we use the empirical distribution function $F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(X_i\leq x)}$, where $x \in \mathbb{R}$ and $\mathbb{I}$ is the indicator function, for a stationary sequence of dependent real-valued random variables $(X_i)_{1\leq i\leq n}$ to estimate the VaR.

The empirical estimator of the VaR $\left(\widehat{VaR}\right)$ (see e.g. Dowd, 2001) is defined by:

$$\widehat{VaR}(q) \equiv F_n^{-1}(q) \equiv \inf\{x : F_n(x) \geq q\}.$$

We note that if we order the independent random variables $X_{n,1} \leq X_{n,2} \leq ... \leq X_{n,n}$ then $\widehat{VaR}_e(q)$ can be written as

$$\widehat{VaR}(q) = X_{n,s}, \qquad s = [nq] + 1.$$

where $[a]$ is the integer part of a.

Next let us recall the definitions of some mixing coefficients which are criteria needed to introduce dependency measures between variables.

Let $(\Omega, \mathcal{K}, P)$ be a probability space and let $\mathcal{A}$, $\mathcal{B}$ be two sub $\sigma-$algebras of $\mathcal{K}$. We define:

1) The $\alpha-$mixing coefficient by:

$$\alpha(\mathcal{A},\mathcal{B}) = \sup_{A\in\mathcal{A},B\in\mathcal{B}} |P(A\cap B) - P(A)P(B)|.$$

2) The $\rho-$mixing coefficient by:

$$\rho(\mathcal{A},\mathcal{B}) = \sup_{f\in L_2(\mathcal{A}),g\in L_2(\mathcal{B})} |corr(f,g)|,$$

where $corr(f,g) = \frac{Cov(f,g)}{\sqrt{Var(f)}\sqrt{Var(g)}}$.

3) The $\varphi-$mixing coefficient by:

$$\varphi(\mathcal{A},\mathcal{B}) = \sup_{A\in\mathcal{A},B\in\mathcal{B}} \left|\frac{P(A\cap B)}{P(A)} - P(B)\right|.$$

Finally, we say that a stationary sequence $(X_i)_{i\in\mathbb{Z}}$ is strong mixing or $\alpha-$mixing, if

$$\alpha_n = \alpha(\sigma(X_i, i\leq 0), \sigma(X_i, i\geq n)) \rightarrow_{n\rightarrow\infty} 0.$$

The paper is organized as follows. The section 2 is related to the notion of oscillation of an empirical process which is defined for each $f_x \in \mathcal{F}$ by:

$$Z_n(f_x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[f_x(X_i) - \mathbb{E}(f_x(X_i))] = \sqrt{n}[F_n(x) - F(x)]$$

where $\mathcal{F}$ is the set of characteristic functions of intervals of the form $(-\infty, x)$ for any $x \in \mathbb{R}$. We study the mean of the modulus of continuity of the empirical process defined by

$$W(n,\delta) := \mathbb{E}\left(\sup_{\|f_x-f_y\|_v\leq\delta}\left|Z_n\left(f_x - f_y\right)\right|\right) \tag{1.2}$$

where $\|f_x\|_v = (\mathbb{E}|f_x|^v)^{\frac{1}{v}}$. Our method is inspired by the work by Ben Hariz (2005) who studied the stochastic equicontinuity of empirical processes indexed by a family of functions.

In the section 3, which is the main part of this work, we prove the consistency as well as a central limit theorem for the $\widehat{VaR}$, i.e.

$$\sqrt{n}\left(\xi_n - \xi\right) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma_\infty^2\left(\xi\right)}{f^2\left(\xi\right)}\right)$$

where

$$\sigma_\infty^2(\xi) = \sum_{i\in\mathbb{Z}} Cov\left(\mathbb{I}_{(X_1\leq\xi)}, \mathbb{I}_{(X_{i+1}\leq\xi)}\right) = Var\left(\mathbb{I}_{(X_1\leq\xi)}\right) + 2\sum_{i=1}^{+\infty} Cov\left(\mathbb{I}_{(X_1\leq\xi)}, \mathbb{I}_{(X_{i+1}\leq\xi)}\right)$$

is assumed to satisfy $0 < \sigma_\infty^2(\xi) < \infty$.

In the section 4, several applications are discussed. Finally, the section 5 is devoted to simulations which illustrate the results.

2. Oscillation of the Empirical Process

First let us introduce the following assumptions:

$H(X)$: $(X_i)_{1\leq i\leq n}$ is a stationary sequence of real-valued random variables with a common distribution function F.

$H(p, X)$: For all positive real numbers $2 \leq v < p < r \leq \infty$ and for any $\varepsilon > 0$, there exists a positive constant $D = D\left(\varepsilon, p, v, r\right) < \infty$ such that for any $f \in \mathcal{F}$

$$\mathbb{E}\left|Z_n\left(f\right)\right|^p \leq D\left(\|f\|_v^p + n^{1+\varepsilon-\frac{p}{2}}\|f\|_r^p\right).$$

$H(F)$: F is continuous in $\mathcal{I} = [\xi - a_n, \xi + a_n]$ where $0 < a_n \rightarrow_{n\rightarrow\infty} 0$, and F has a density function f which is continuous and $0 < f\left(\xi\right) < \infty$.

For $0 < b_n \rightarrow_{n\rightarrow\infty} 0$ we denote,

$$a_n \ll b_n \Leftrightarrow \left\{a_n < b_n \text{ and } \frac{a_n}{b_n} \rightarrow_{n\rightarrow\infty} 0\right\}.$$

In the proofs C denote constant where values may change from one line to another. We will now focus on the modulus of continuity of an empirical process $(X_i)_{1\leq i\leq n}$.

Theorem 1 *Under conditions $H(X)$ and $H(p, X)$, there exists $C = C\left(\varepsilon, p, v, r\right) < \infty$ such that for $\delta > n^{\frac{\frac{1+\varepsilon}{p}-1}{v\left(1+\frac{1}{p}-\frac{1}{r}\right)}}$,*

$$W\left(n, \delta\right) \leq C\cdot\left(n^{-\frac{1}{2}+\frac{2+\varepsilon-\frac{p}{r}}{p+1-\frac{p}{r}}} + \delta^{\left(1-\frac{v}{p}\right)}\right).$$

If in addition $\varepsilon < \frac{p}{2}\left(1 - \frac{1}{p} + \frac{1}{r}\right) - 1$ and $\delta = \delta_n \rightarrow 0$, then

$$\lim_{n\rightarrow\infty} W\left(n, \delta_n\right) = 0.$$

Remark

- When $r = p$ the result becomes for $\delta > n^{\frac{1+\varepsilon-p}{vp}}$,

$$W\left(n, \delta\right) \leq C\cdot\left(\ln n \cdot n^{-\frac{1}{2}+\frac{1+\varepsilon}{p}} + \delta^{\left(1-\frac{v}{p}\right)}\right). \tag{2.1}$$

- If F is L–Lipschitz, then for $\delta_0 > \frac{1}{C(v,L)} n^{\frac{\frac{1+\varepsilon}{p}-1}{1+\frac{1}{p}-\frac{1}{r}}}$,

$$\mathbb{E}\left[\sup_{|x-y|\leq\delta_0}\left|Z_n\left(f_x - f_y\right)\right|\right] \leq C\cdot\left(n^{-\frac{1}{2}+\frac{2+\varepsilon-\frac{p}{r}}{p+1-\frac{p}{r}}} + C\left(v, p, L\right)\cdot\delta_0^{\left(\frac{1}{v}-\frac{1}{p}\right)}\right).$$

Proof of Theorem 1. Let $N\left(k\right) = N_{[.]}\left(2^{-k}, \|.\|_v, \mathcal{F}\right), k \in \mathbb{N}$ (the bracketing number) be the minimal number of brackets which are of a norm $\|.\|_v$ less than or equal 2^{-k} needed to cover $\mathcal{F}$. As $N\left(k\right) \leq 2.2^{vk}$ is finite (see e.g. Van der Vaart & Wellner, 1996, ex 2.5.4 in p. 129), there exists a finite sequence

$$\{f_{x_k(i)}, \Delta_{x_k(i)} = \mathbb{I}_{(x_k(i)\leq.\leq x_k(i+1))}\}_{1\leq i\leq 2^{vk}}$$

such that:

1) $\left\|\Delta_{x_k(i)}\right\|_v \le 2^{-k}$,

2) $\forall\, f_x \in \mathcal{F}, \exists\, i : \left|f_x - f_{x_k(i)}\right| \le \Delta_{x_k(i)}$.

We set $(\pi_k(f), \Delta_k(f))$ the first pair $(f_{x_k(i)}, \Delta_{x_k(i)})$ which satisfies $\left|f_x - f_{x_k(i)}\right| \le \Delta_{x_k(i)}$. Let q_0, k and $q_1 \in \mathbb{N}$ such that $q_0 \le k \le q_1$, we define for $1 \le i \le 2^{vq_0}$,

$$E_i = \left\{f \in \mathcal{F} : \pi_{q_0}(f) = f_{x_{q_0}(i)}\right\},$$

then the sets E_i form a partition of $\mathcal{F}$. For $\delta \sim 2^{-q_0} \Leftrightarrow q_0 \sim -\frac{\ln\delta}{\ln 2}$, we define:

$$F_{i,j} = \left\{\left(f_x, f_y\right) \in \mathcal{F}\times\mathcal{F} : f_x \in E_i, f_y \in E_j, \left\|f_x - f_y\right\|_v \le \delta\right\}.$$

Let now $\Lambda = \left\{(i,j) : F_{i,j} \ne \emptyset\right\}$. For every pair $(i,j) \in \Lambda$, we fix an element of $F_{i,j}$ and denote this pair $\left(\phi_{i,j}, \psi_{i,j}\right)$. Let $\left(f_x, f_y\right)$ be a pair satisfying $\left\|f_x - f_y\right\|_v \le \delta$, then $\left(f_x, f_y\right) \in F_{i,j}$ for some $(i,j) \in \Lambda$. We write

$$f_x - f_y = f_x - \pi_{q_0}(f_x) + \pi_{q_0}(f_x) - \phi_{i,j} + \phi_{i,j} - \psi_{i,j} + \psi_{i,j} - \pi_{q_0}\left(f_y\right) + \pi_{q_0}\left(f_y\right) - f_y$$

but $\pi_{q_0}(f_x) = \pi_{q_0}\left(\phi_{i,j}\right)$ and $\pi_{q_0}\left(f_y\right) = \pi_{q_0}\left(\psi_{i,j}\right)$ since $f_x, \phi_{i,j} \in E_i$, $f_y, \psi_{i,j} \in E_j$. Consequently:

$$\sup_{\left\|f_x - f_y\right\|_v \le \delta} \left|Z_n\left(f_x - f_y\right)\right| \le 4 \sup_{f_x \in \mathcal{F}} \left|Z_n\left(f_x - \pi_{q_0}(f_x)\right)\right| + \sup_{(i,j)\in\Lambda} \left|Z_n\left(\phi_{i,j} - \psi_{i,j}\right)\right|$$

That gives by applying the expectation:

$$\begin{aligned}
\mathbb{E}\left(\sup_{\left\|f_x - f_y\right\|_v \le \delta} \left|Z_n\left(f_x - f_y\right)\right|\right) &\le 4\mathbb{E}\left(\sup_{f_x \in \mathcal{F}} \left|Z_n\left(f_x - \pi_{q_0}(f_x)\right)\right|\right) + \mathbb{E}\left(\sup_{(i,j)\in\Lambda} \left|Z_n\left(\phi_{i,j} - \psi_{i,j}\right)\right|\right) \\
&\equiv 4E_1 + E_2.
\end{aligned}$$

In order to control the terms E_1 and E_2, we put $\|Z_n(f)\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |Z_n(f)|$, and we use the following inequality due to Pisier: For all random variables $Z_1, Z_2, ..., Z_N$

$$\mathbb{E}\left[\max_{1\le i\le N} |Z_i|\right] \le N^{\frac{1}{p}} \max_{1\le i\le N} \left(\mathbb{E}|Z_i|^p\right)^{\frac{1}{p}}.$$

<u>**Control of E_1**</u>: For $f \in \mathcal{F}$, we write:

$$f - \pi_{q_0}(f) = f - \pi_{q_1}(f) + \sum_{k=q_0+1}^{q_1} \left[\pi_k(f) - \pi_{k-1}(f)\right].$$

Therefore,

$$\begin{aligned}
E_1 &\equiv \mathbb{E}\left\|Z_n\left(f - \pi_{q_0}(f)\right)\right\|_{\mathcal{F}} \\
&\le \mathbb{E}\left\|Z_n\left(f - \pi_{q_1}(f)\right)\right\|_{\mathcal{F}} + \sum_{k=q_0+1}^{q_1} \mathbb{E}\left\|Z_n\left(\pi_k(f) - \pi_{k-1}(f)\right)\right\|_{\mathcal{F}} \\
&\le E_{1,q_1+1} + 2\sqrt{n} \sup_{f\in\mathcal{F}} \mathbb{E}\left|\Delta_{q_1}(f)\right| + \sum_{k=q_0+1}^{q_1} E_{1,k}
\end{aligned}$$

where $E_{1,k} = \mathbb{E}\left\|Z_n\left(\pi_k(f) - \pi_{k-1}(f)\right)\right\|_{\mathcal{F}}$, $q_0 + 1 \le k \le q_1$ and $E_{1,q_1+1} = \mathbb{E}\left\|Z_n\left(\Delta_{q_1}(f)\right)\right\|_{\mathcal{F}}$. Note that $\pi_k(f) - \pi_{k-1}(f) = \pi_k(f) - \pi_{k-1}\left(\pi_k(f)\right)$ and $\pi_k(f)$ take values on a finite set $N(k) \le 2.2^{vk}$. Then using Pisier's inequality, we can write:

$$E_{1,k} \le 2^{\frac{vk}{p}} \max_{g\in\pi_k(\mathcal{F})} \left\|Z_n\left(g - \pi_{k-1}(g)\right)\right\|_p.$$

Apply $H(p,X)$ to $h = g - \pi_{k-1}(g)$ to get:

$$\begin{aligned}\|Z_n(h)\|_p &\leq D^{\frac{1}{p}}\left(\|h\|_v^p + n^{1+\varepsilon-\frac{p}{2}}\|h\|_r^p\right)^{\frac{1}{p}}\\ &\leq D^{\frac{1}{p}}\left(\|h\|_v + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}\|h\|_r\right)\end{aligned}$$

Using the fact that

$$\|X\|_r \leq \|X\|_v^{\frac{v}{r}} \times \|X\|_\infty^{\frac{r-v}{r}},$$

we obtain

$$\begin{aligned}\|Z_n(h)\|_p &\leq D^{\frac{1}{p}}\cdot\left(2^{-(k-1)} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}2^{-\frac{(k-1)v}{r}}\right)\\ &\leq 2D^{\frac{1}{p}}\cdot\left(2^{-k} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}2^{-\frac{kv}{r}}\right).\end{aligned}$$

Hence,

$$\begin{aligned}E_{1,k} &\leq 2D^{\frac{1}{p}}\cdot 2^{\frac{vk}{p}}\left(2^{-k} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}2^{-\frac{kv}{r}}\right)\\ &\leq 2D^{\frac{1}{p}}\cdot\left(2^{-k\left(1-\frac{v}{p}\right)} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}2^{k\left(\frac{v}{p}-\frac{v}{r}\right)}\right)\\ &\leq C\cdot\left(2^{-k\left(1-\frac{v}{p}\right)} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}2^{k\left(\frac{v}{p}-\frac{v}{r}\right)}\right)\end{aligned}$$

Similarly for E_{1,q_1+1}:

$$E_{1,q_1+1} \leq C\cdot\left(2^{-(q_1+1)\left(1-\frac{v}{p}\right)} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}2^{(q_1+1)\left(\frac{v}{p}-\frac{v}{r}\right)}\right).$$

Finally, using that $\mathbb{E}\left|\Delta_{q_1}(f)\right| = \left\|\Delta_{q_1}(f)\right\|_v^v \leq 2^{-q_1 v}$, we obtain:

$$E_1 \leq C\cdot\sqrt{n}2^{-q_1 v} + \sum_{k=q_0+1}^{q_1+1} E_{1,k} \leq C\cdot\sqrt{n}2^{-q_1 v} + C\cdot\sum_{k=q_0+1}^{q_1+1}\left(2^{-k\left(1-\frac{v}{p}\right)} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}2^{k\left(\frac{v}{p}-\frac{v}{r}\right)}\right) \tag{2.2}$$

$$\leq C\cdot\left(\sqrt{n}2^{-q_1 v} + 2^{-q_0\left(1-\frac{v}{p}\right)} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}\left[2^{q_1\left(\frac{v}{p}-\frac{v}{r}\right)} - 2^{q_0\left(\frac{v}{p}-\frac{v}{r}\right)}\right]\right).$$

Then,

$$E_1 \leq C\cdot\left(\sqrt{n}2^{-q_1 v} + 2^{-q_0\left(1-\frac{v}{p}\right)} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}\left[2^{q_1\left(\frac{v}{p}-\frac{v}{r}\right)} - 2^{q_0\left(\frac{v}{p}-\frac{v}{r}\right)}\right]\right). \tag{2.3}$$

Control of E_2: Noting that $|\Lambda| \leq 2\times 2^{vq_0}$ (since if $F_{i,j} \neq \phi$, then $j = \{i-1, i, i+1\}$, because $\left|f_x - f_{x_{q_0}(i)}\right| \leq \Delta_{x_{q_0}(i)}$ and $\left\|\Delta_{x_{q_0}(i)}\right\|_v \leq 2^{-q_0}$) and $\left\|\phi_{i,j} - \psi_{i,j}\right\|_v \leq \delta$, using the inequality of Pisier, we get

$$\begin{aligned}E_2 &= \mathbb{E}\left(\sup_{(i,j)\in\Lambda}\left|Z_n\left(\phi_{i,j} - \psi_{i,j}\right)\right|\right)\\ &\leq 2^{\frac{vq_0}{p}}\max_{(i,j)\in\Lambda}\left\|Z_n\left(\phi_{i,j} - \psi_{i,j}\right)\right\|_p.\end{aligned}$$

Again by $H(p,X)$,

$$\begin{aligned}E_2 &\leq 2^{\frac{vq_0}{p}}\left[D^{\frac{1}{p}}\cdot\left(\left\|\phi_{i,j} - \psi_{i,j}\right\|_v + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}\left\|\phi_{i,j} - \psi_{i,j}\right\|_r\right)\right]\\ &\leq 2^{\frac{vq_0}{p}}D^{\frac{1}{p}}\cdot\left(\delta + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}\delta^{\frac{v}{r}}\right)\end{aligned}$$

Then,

$$E_2 \leq D^{\frac{1}{p}}\cdot 2^{\frac{vq_0}{p}}\left(\delta + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}\delta^{\frac{v}{r}}\right). \tag{2.4}$$

Thus, from (2.3) and (2.4) we conclude that:

$$W(n,\delta) \leq C\cdot\left[\sqrt{n}2^{-q_1 v} + 2^{-q_0\left(1-\frac{v}{p}\right)} + 2^{\frac{vq_0}{p}}\delta + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}}\left(2^{q_1\left(\frac{v}{p}-\frac{v}{r}\right)} - 2^{q_0\left(\frac{v}{p}-\frac{v}{r}\right)} + 2^{\frac{vq_0}{p}}\delta^{\frac{v}{r}}\right)\right].$$

We have $\delta \sim 2^{-q_0}$ then $2^{\frac{vq_0}{p}} \cdot \delta \sim 2^{-q_0\left(1-\frac{v}{p}\right)} \sim \delta^{\left(1-\frac{v}{p}\right)}$, hence

$$W(n,\delta) \leq C \cdot \left[\sqrt{n}2^{-q_1 v} + \delta^{\left(1-\frac{v}{p}\right)} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}} 2^{q_1\left(\frac{v}{p}-\frac{v}{r}\right)} \right].$$

Take q_1 such that $\sqrt{n}2^{-q_1 v} \sim n^{\frac{1+\varepsilon}{p}-\frac{1}{2}} 2^{q_1\left(\frac{v}{p}-\frac{v}{r}\right)}$ then

$$2^{q_1} \sim n^{\frac{1-\frac{1+\varepsilon}{p}}{v\left(1+\frac{1}{p}-\frac{1}{r}\right)}} \Rightarrow q_1 \sim \frac{\left(1-\frac{1+\varepsilon}{p}\right)\ln n}{v\left(1+\frac{1}{p}-\frac{1}{r}\right)\ln 2}.$$

Therefore,

$$W(n,\delta) \leq C \cdot \left(n^{-\frac{1}{2}+\frac{2+\varepsilon-\frac{p}{r}}{p+1-\frac{p}{r}}} + \delta^{\left(1-\frac{v}{p}\right)} \right).$$

As q_1 and q_0 have to satisfy $q_0 < q_1$ then $\delta > n^{\frac{\frac{1+\varepsilon}{p}-1}{v\left(1+\frac{1}{p}-\frac{1}{r}\right)}}$. And to ensure that $W(n,\delta) \to_{\{n\to\infty,\delta\to 0\}} 0$, we need $-\frac{1}{2} + \frac{2+\varepsilon-\frac{p}{r}}{p+1-\frac{p}{r}} < 0$ which is this

$$\varepsilon < \frac{p}{2}\left(1-\frac{1}{p}+\frac{1}{r}\right)-1.$$

Proof of Remark 1. The proof of the first point of the Remark 1 has the same steps of the proof of Theorem 1 up to the inequality (2.2). This relation becomes in the case where $r = p$,

$$\begin{aligned} E_1 &\leq C \cdot \sqrt{n}2^{-q_1 v} + C \cdot \sum_{k=q_0+1}^{q_1+1} \left(2^{-k\left(1-\frac{v}{p}\right)} + n^{\frac{1+\varepsilon}{p}-\frac{1}{2}} \right) \\ &\leq C \cdot \left(\sqrt{n}2^{-q_1 v} + 2^{-q_0\left(1-\frac{v}{p}\right)} + q_1 n^{\frac{1+\varepsilon}{p}-\frac{1}{2}} \right). \end{aligned}$$

Since

$$q_1 \sim \frac{1}{v\ln 2}\left(1-\frac{1+\varepsilon}{p}\right)\ln n.$$

Therefore,

$$W(n,\delta) \leq C \cdot \left(\ln n . n^{-\frac{1}{2}+\frac{1+\varepsilon}{p}} + \delta^{\left(1-\frac{v}{p}\right)} \right).$$

3. Limit Theorems for the Empirical VaR

In this part we will apply the results of the previous section on the fluctuations of the empirical process to deduce asymptotic results on the $\widehat{VaR}(q)$.

Theorem 2 *Under conditions $H(X)$, $H(F)$ and $H(p,X)$ where $\varepsilon < \frac{p}{2}\left(1-\frac{1}{p}+\frac{1}{r}\right)-1$, we have for $a_n \gg n^{-\frac{1}{2}}$,*

$$|\xi_n - \xi| = o_p(a_n).$$

If in addition

$$\sqrt{n}\left(F_n(\xi) - F(\xi)\right) \to^d \mathcal{N}\left(0, \sigma^2_\infty(\xi)\right),$$

then

$$\sqrt{n}(\xi_n - \xi) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma^2_\infty(\xi)}{f^2(\xi)}\right).$$

The proof of the previous theorem is based on the two following lemmas:

Lemma 1 *Under conditions $H(X)$, $H(F)$ and $H(p,X)$ where $\varepsilon \leq \frac{p}{2}-1$, we have for $a_n > 0$,*

$$\mathbb{P}\left(|\xi_n - \xi| > a_n\right) \leq C(\varepsilon, p, v, r, \xi) . \left(n^{\frac{1}{2}} a_n\right)^{-p}.$$

If in addition $a_n \gg n^{-\frac{1}{2}}$, then

$$|\xi_n - \xi| = o_p(a_n).$$

Proof of Lemma 1. Let $s = [nq] + 1$. Then, we note that

$$\begin{aligned}\mathbb{P}(\xi_n < \xi - a_n) &= \mathbb{P}(s \text{ or more of the } X_i\,(1 \leq i \leq n) \text{ are } < \xi - a_n)\\ &= \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{I}_{(X_i<\xi-a_n)} \geq s\right)\\ &= \mathbb{P}\left(F_n(\xi - a_n) \geq \frac{s}{n}\right)\\ &= \mathbb{P}\left(F_n(\xi - a_n) - F(\xi - a_n) \geq \frac{s}{n} - F(\xi - a_n)\right).\end{aligned}$$

Since

$$F_n(\xi_n) = \frac{s}{n} = F(\xi) + O\left(n^{-1}\right), \text{ (see e.g. Sen, 1972)}$$

then, using $H(F)$ and the first-order Taylor expansion of $F(\xi - a_n)$, one obtains

$$\frac{s}{n} - F(\xi - a_n) = f(\xi)\, a_n\,[1 + o(1)].$$

Then

$$\mathbb{P}(\xi_n < \xi - a_n) = \mathbb{P}(F_n(\xi - a_n) - F(\xi - a_n) \geq f(\xi)\, a_n\,[1 + o(1)]).$$

And by Markov's inequality, this is bounded by

$$\begin{aligned}\mathbb{P}(\xi_n < \xi - a_n) &\leq \left(\frac{1}{f(\xi)\, a_n\,[1 + o(1)]}\right)^p \mathbb{E}\,[F_n(\xi - a_n) - F(\xi - a_n)]^p,\\ &\leq C \cdot \left(\frac{1}{f(\xi)\, a_n}\right)^p \mathbb{E}\,|F_n(\xi - a_n) - F(\xi - a_n)|^p.\end{aligned}$$

But,

$$\mathbb{E}\,|F_n(\xi - a_n) - F(\xi - a_n)|^p = \left(\frac{1}{\sqrt{n}}\right)^p \mathbb{E}\left|Z_n\left(f_{(\xi-a_n)}\right)\right|^p.$$

By $H(p, X)$,

$$\begin{aligned}\mathbb{E}\,|F_n(\xi - a_n) - F(\xi - a_n)|^p &\leq \left(\frac{1}{\sqrt{n}}\right)^p D \cdot \left(\left\|\mathbb{I}_{(X_i<\xi-a_n)}\right\|_v^p + n^{1+\varepsilon-\frac{p}{2}} \cdot \left\|\mathbb{I}_{(X_i<\xi-a_n)}\right\|_r^p\right)\\ &\leq n^{\frac{-p}{2}} D \cdot \left(F(\xi - a_n)^{\frac{p}{v}} + n^{1+\varepsilon-\frac{p}{2}} F(\xi - a_n)^{\frac{p}{r}}\right).\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{P}(\xi_n < \xi - a_n) &\leq C \cdot \left(\frac{1}{f(\xi)\, a_n}\right)^p n^{\frac{-p}{2}} D \cdot \left(F(\xi - a_n)^{\frac{p}{v}} + n^{1+\varepsilon-\frac{p}{2}} F(\xi - a_n)^{\frac{p}{r}}\right)\\ &\leq C \cdot D\left(\frac{1}{f(\xi)}\right)^p n^{\frac{-p}{2}} a_n^{-p} \left(F(\xi - a_n)^{\frac{p}{v}} + n^{1+\varepsilon-\frac{p}{2}} F(\xi - a_n)^{\frac{p}{r}}\right)\\ &\leq C(\varepsilon, p, v, r, \xi) \cdot \left(1 + n^{1+\varepsilon-\frac{p}{2}}\right) n^{\frac{-p}{2}} a_n^{-p}.\end{aligned}$$

Consequently for $0 < a_n$ and $\varepsilon \leq \frac{p}{2} - 1$

$$\mathbb{P}(\xi_n < \xi - a_n) \leq C(\varepsilon, p, v, r, \xi) \cdot n^{\frac{-p}{2}} a_n^{-p}. \tag{3.1}$$

For the second term, we note that:

$$\begin{aligned}\mathbb{P}(\xi_n > \xi + a_n) &= \mathbb{P}(s \text{ or less of the } X_i\,(1 \leq i \leq n) \text{ are } < \xi + a_n)\\ &= \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{I}_{(X_i<\xi+a_n)} \leq s\right)\\ &= \mathbb{P}\left(F_n(\xi + a_n) \leq \frac{s}{n}\right)\\ &= \mathbb{P}\left(F_n(\xi + a_n) - F(\xi + a_n) \leq \frac{s}{n} - F(\xi + a_n)\right).\end{aligned}$$

But, using $H(F)$ and the first-order Taylor expansion of $F(\xi + a_n)$, one obtains

$$\begin{aligned}\mathbb{P}(\xi_n > \xi + a_n) &= \mathbb{P}(F_n(\xi + a_n) - F(\xi + a_n) \leq -f(\xi)\, a_n [1 + o(1)]) \\ &= \mathbb{P}(F(\xi + a_n) - F_n(\xi + a_n) \geq f(\xi)\, a_n [1 + o(1)]).\end{aligned}$$

and by Markov's inequality, this is bounded by

$$\begin{aligned}\mathbb{P}(\xi_n > \xi + a_n) &\leq \left(\frac{1}{f(\xi)\, a_n [1 + o(1)]}\right)^p \mathbb{E}[F(\xi + a_n) - F_n(\xi + a_n)]^p, \\ &\leq C \cdot \left(\frac{1}{f(\xi)\, a_n}\right)^p \mathbb{E}|F_n(\xi + a_n) - F(\xi + a_n)|^p.\end{aligned}$$

In the same way for the first term, we have

$$\mathbb{E}|F_n(\xi + a_n) - F(\xi + a_n)|^p = \left(\frac{1}{\sqrt{n}}\right)^p \mathbb{E}\left|Z_n\left(f_{(\xi + a_n)}\right)\right|^p.$$

By $H(p, X)$,

$$\begin{aligned}\mathbb{E}|F_n(\xi + a_n) - F(\xi + a_n)|^p &\leq \left(\frac{1}{\sqrt{n}}\right)^p D \cdot \left(\left\|\mathbb{I}_{(X_i < \xi + a_n)}\right\|_v^p + n^{1+\varepsilon-\frac{p}{2}} \cdot \left\|\mathbb{I}_{(X_i < \xi + a_n)}\right\|_r^p\right) \\ &\leq n^{\frac{-p}{2}} D. \left(F(\xi + a_n)^{\frac{p}{v}} + n^{1+\varepsilon-\frac{p}{2}} F(\xi + a_n)^{\frac{p}{r}}\right).\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{P}(\xi_n > \xi + a_n) &\leq C \cdot \left(\frac{1}{f(\xi)\, a_n}\right)^p n^{\frac{-p}{2}} D \cdot \left(F(\xi + a_n)^{\frac{p}{v}} + n^{1+\varepsilon-\frac{p}{2}} F(\xi + a_n)^{\frac{p}{r}}\right) \\ &\leq C \cdot D \cdot \left(\frac{1}{f(\xi)}\right)^p n^{\frac{-p}{2}} a_n^{-p} \left(F(\xi + a_n)^{\frac{p}{v}} + n^{1+\varepsilon-\frac{p}{2}} F(\xi + a_n)^{\frac{p}{r}}\right) \\ &\leq C(\varepsilon, p, v, r, \xi) \cdot \left(1 + n^{1+\varepsilon-\frac{p}{2}}\right) n^{\frac{-p}{2}} a_n^{-p}.\end{aligned}$$

Consequently for $0 < a_n$ and $\varepsilon \leq \frac{p}{2} - 1$

$$\mathbb{P}(\xi_n > \xi + a_n) \leq C(\varepsilon, p, v, r, \xi) \cdot n^{\frac{-p}{2}} a_n^{-p}. \tag{3.2}$$

Thus, from (3.1) and (3.2) we conclude for $0 < a_n$ and $\varepsilon \leq \frac{p}{2} - 1$

$$\mathbb{P}(|\xi_n - \xi| > a_n) \leq C(\varepsilon, p, v, r, \xi) \cdot \left(n^{\frac{1}{2}} a_n\right)^{-p}.$$

Finally, if $a_n \gg n^{-\frac{1}{2}}$, then

$$\mathbb{P}(|\xi_n - \xi| > a_n) \to_{n\to\infty} 0.$$

The following lemma studies the proximity between $Z_n\left(f_{\xi_n}\right) = \sqrt{n}(F_n(\xi_n) - F(\xi_n))$ and $Z_n\left(f_\xi\right) = \sqrt{n}(F_n(\xi) - F(\xi))$.

Lemma 2 *Under conditions $H(X)$, $H(F)$ and $H(p, X)$ where $\varepsilon < \frac{p}{2}\left(1 - \frac{1}{p} + \frac{1}{r}\right) - 1$, we have for $a_n \gg n^{-\frac{1}{2}}$ and $b_n \gg \max\left(n^{-\frac{1}{2}+\frac{2+\varepsilon-\frac{p}{r}}{p+1-\frac{p}{r}}}, a_n^{\left(\frac{1}{v}-\frac{1}{p}\right)}\right)$,*

$$\left|\sqrt{n}(F_n(\xi_n) - F(\xi_n)) - \sqrt{n}(F_n(\xi) - F(\xi))\right| = o_p(b_n).$$

Proof of Lemma 2. Let $0 < a_n$ and $0 < b_n$, we note that

$$\begin{aligned}\mathbb{P}\left(\left|Z_n\left(f_{\xi_n} - f_\xi\right)\right| > b_n\right) &= \mathbb{P}\left(\left|Z_n\left(f_{\xi_n} - f_\xi\right)\right| > b_n \cap |\xi_n - \xi| \leq a_n\right) + \mathbb{P}\left(\left|Z_n\left(f_{\xi_n} - f_\xi\right)\right| > b_n \cap |\xi_n - \xi| > a_n\right) \\ &\leq \mathbb{P}\left(\left|Z_n\left(f_{\xi_n} - f_\xi\right)\right| \mathbb{I}_{(|\xi_n - \xi| \leq a_n)} > b_n\right) + \mathbb{P}(|\xi_n - \xi| > a_n).\end{aligned}$$

If $H(p, X)$ is verified for $\varepsilon < \frac{p}{2}\left(1 - \frac{1}{p} + \frac{1}{r}\right) - 1 \leq \frac{p}{2} - 1$ and $0 < a_n$, then by Lemma 1:

$$\mathbb{P}\left(|\xi_n - \xi| > a_n\right) \leq C \cdot \left(n^{\frac{1}{2}} a_n\right)^{-p}.$$

If $H(F)$ is verified, then F is locally Lipschitz, then for $|y - \xi| \leq a_n$, we have

$$\left\|f_y - f_\xi\right\|_v = |F(y) - F(\xi)|^{\frac{1}{v}} \leq C(v, \xi) \cdot |y - \xi|^{\frac{1}{v}} \leq C(v, \xi) \cdot a_n^{\frac{1}{v}}.$$

In addition, by Markov's inequality and Theorem 1 for $a_n > \frac{1}{C(v,\xi)} n^{\frac{\frac{1+\varepsilon}{p} - 1}{1 + \frac{1}{p} - \frac{1}{r}}}$

$$\begin{aligned} \mathbb{P}\left(\left|Z_n\left(f_{\xi_n} - f_\xi\right)\right| \mathbb{I}_{(|\xi_n - \xi| \leq a_n)} > b_n\right) &\leq \frac{1}{b_n} \mathbb{E}\left|\sup_{|y-\xi| \leq a_n} \left|Z_n\left(f_y - f_\xi\right)\right|\right| \\ &\leq C \cdot b_n^{-1}\left(n^{-\frac{1}{2} + \frac{2+\varepsilon-\frac{p}{r}}{p+1-\frac{p}{r}}} + C.a_n^{\left(\frac{1}{v} - \frac{1}{p}\right)}\right). \end{aligned}$$

Consequently, for $a_n > \frac{1}{C(v,\xi)} n^{\frac{\frac{1+\varepsilon}{p} - 1}{1 + \frac{1}{p} - \frac{1}{r}}}$ where $\varepsilon < \frac{p}{2}\left(1 - \frac{1}{p} + \frac{1}{r}\right) - 1$ and $0 < b_n$

$$\mathbb{P}\left(\left|Z_n\left(f_{\xi_n} - f_\xi\right)\right| > b_n\right) \leq C \cdot \left[b_n^{-1}\left(n^{-\frac{1}{2} + \frac{2+\varepsilon-\frac{p}{r}}{p+1-\frac{p}{r}}} + C \cdot a_n^{\left(\frac{1}{v} - \frac{1}{p}\right)}\right) + \left(n^{\frac{1}{2}} a_n\right)^{-p}\right].$$

If $a_n \gg n^{-\frac{1}{2}}$ and $b_n \gg \max\left(n^{-\frac{1}{2} + \frac{2+\varepsilon-\frac{p}{r}}{p+1-\frac{p}{r}}}, a_n^{\left(\frac{1}{v} - \frac{1}{p}\right)}\right)$, then

$$\left|Z_n\left(f_{\xi_n} - f_\xi\right)\right| = o_p(b_n).$$

Finally, by the definition of $Z_n(f_x)$, we obtain

$$\left|\sqrt{n}\left(F_n(\xi_n) - F(\xi_n)\right) - \sqrt{n}\left(F_n(\xi) - F(\xi)\right)\right| = o_p(b_n).$$

Proof of Theorem 2. By Lemmas 1 and 2 for $a_n \gg n^{-\frac{1}{2}}$ and $b_n \gg \max\left(n^{-\frac{1}{2} + \frac{2+\varepsilon-\frac{p}{r}}{p+1-\frac{p}{r}}}, a_n^{\left(\frac{1}{v} - \frac{1}{p}\right)}\right)$, we have

$$\left|\sqrt{n}\left(F_n(\xi_n) - F(\xi_n)\right) - \sqrt{n}\left(F_n(\xi) - F(\xi)\right)\right| = o_p(b_n). \tag{3.3}$$

Since

$$F_n(\xi_n) = \frac{s}{n} = F(\xi) + O\left(n^{-1}\right),$$

then,

$$\sqrt{n}\left(F_n(\xi_n) - F(\xi_n)\right) = \sqrt{n}\left(F(\xi) - F(\xi_n)\right) + O\left(n^{-\frac{1}{2}}\right). \tag{3.4}$$

If $H(F)$ is satisfied, then by the Mean Value Theorem of $F(\xi) - F(\xi_n)$,

$$F(\xi) - F(\xi_n) = (\xi - \xi_n) f(\theta\xi_n + (1 - \theta)\xi)$$

where $\theta \in [0, 1]$. Then

$$\sqrt{n}\left(F_n(\xi_n) - F(\xi_n)\right) = \sqrt{n}(\xi - \xi_n) f(\theta\xi_n + (1 - \theta)\xi) + O\left(n^{-\frac{1}{2}}\right).$$

Hence,

$$\left|\sqrt{n}(\xi - \xi_n) f(\theta\xi_n + (1 - \theta)\xi) + O\left(n^{-\frac{1}{2}}\right) - \sqrt{n}\left(F_n(\xi) - F(\xi)\right)\right| = o_p(b_n). \tag{3.5}$$

But we have,

$$\sqrt{n}\left(F_n(\xi) - F(\xi)\right) \to^d \mathcal{N}\left(0, \sigma_\infty^2(\xi)\right).$$

And by Lemma 1 for $a_n \gg n^{-\frac{1}{2}}$,

$$f(\theta\xi_n + (1 - \theta)\xi) = \left[f\left(\xi + o_p(a_n)\right)\right] \to_{n\to\infty} f(\xi) \text{ in probability.} \tag{3.6}$$

Then by (3.3), (3.4), (3.5), (3.6) and Slutsky's Theorem (Cramér, 1946, p. 254), we have:

$$\sqrt{n}\left(f(\xi)(\xi - \xi_n)\right) \to^d \mathcal{N}\left(0, \sigma^2_\infty(\xi)\right).$$

Which is equivalent in the result to,

$$\sqrt{n}(\xi_n - \xi) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma^2_\infty(\xi)}{f^2(\xi)}\right).$$

4. Applications

In this section we apply the previous results for different sequences. Using the findings of Hu (2003, p. 1124) and Peligrad (1985, Theorem 2.1, p. 1305), we apply our result to $\varphi-$mixing case. Making use of the result of Utev and Peligrad (2003, Theorem 2.1 and 2.2), we apply our result to the $\rho-$mixing case and to $\alpha-$mixing by mean of the results in Shao and Yu (1996, Theorem 4.1) and Rio (1997, Theorem 7.2). We also consider the nonlinear functional of Gaussian sequences to which we apply the result of Ben Hariz (2011) and Breuer and Major (1983). Finally we compare the results with those in the existing literature.

4.1 $\varphi-$mixing Process

Corollary 1 *Under condition $H(X)$, if the $\varphi-$mixing coefficient satisfies*

$$\sum_{i=0}^{\infty} \varphi^{\frac{1}{p}}\left(2^i\right) < \infty \quad \textit{with } p > 2,$$

Then, for $\delta > n^{-\frac{1}{2}\left(1-\frac{1}{p}\right)}$, there is a positive constant $C(p, \varphi(.))$ such that for any $f \in \mathcal{F}$

$$\mathbb{E}\left[\sup_{\|f_x - f_y\|_2 \le \delta} \left|Z_n\left(f_x - f_y\right)\right|\right] \le C(p, \varphi(.)) \cdot \left(\ln n.n^{\frac{1}{p}-\frac{1}{2}} + \delta^{\left(1-\frac{2}{p}\right)}\right).$$

If $H(F)$ is verified, then for $a_n \gg n^{-\frac{1}{2}}$ we have

$$|\xi_n - \xi| = o_p(a_n).$$

and if in addition $0 < \sigma^2_\infty < \infty$, then

$$\sqrt{n}(\xi_n - \xi) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma^2_\infty(\xi)}{f^2(\xi)}\right).$$

Proof of Corollary 1. When $(X_i)_{i\ge 1}$ are identically distributed, using a Lemma by Hu (2003, p. 1124), if

$$\sum_{i=0}^{\infty} \varphi^{\frac{1}{p}}\left(2^i\right) < \infty,$$

then, there exists a positive constant $K = K(p, \varphi(.))$ such that for all $n \ge 1$ and for any f

$$\mathbb{E}\left|Z_n(f)\right|^p \le C\left(p, \varphi(\cdot)\right).\left(\|f\|_2^p + n^{1-\frac{p}{2}}\|f\|_p^p\right).$$

Then $H(p, X)$ is satisfied with $\varepsilon = 0, v = 2$ and $p = r$. Apply now Theorem 1 for $\delta > n^{-\frac{1}{2}\left(1-\frac{1}{p}\right)}$, to obtain

$$\mathbb{E}\left[\sup_{\|f_x - f_y\|_2 \le \delta} \left|Z_n\left(f_x - f_y\right)\right|\right] \le C \cdot \left(\ln n \cdot n^{\frac{1}{p}-\frac{1}{2}} + \delta^{\left(1-\frac{2}{p}\right)}\right).$$

If $H(F)$ is verified and $a_n \gg n^{-\frac{1}{2}}$, then by Lemma 1 for $p > 2$ we obtain

$$|\xi_n - \xi| = o_p(a_n).$$

To show that

$$\sqrt{n}\left(F_n(\xi) - F(\xi)\right) \to^d \mathcal{N}\left(0, \sigma^2_\infty(\xi)\right),$$

we will apply a result by Peligrad (1985, Theorem 2.1, p. 1305) with $Y_i \equiv \mathbb{I}_{(X_i \le \xi)} - F(\xi)$, $\sigma_n^2 = \mathbb{E}\left[\sum_{i=1}^n Y_i\right]^2$ and $W_n(t) := \frac{1}{\sigma_n}\sum_{i=1}^{[nt]} Y_i$, $t \in [0,1]$ and $0 < \sigma_\infty^2 < \infty$. If we have $0 < \sigma_\infty^2 < \infty$, then $\frac{\sigma_n^2}{n} \to_{n\to\infty} \sigma_\infty^2$.

The condition (L) therein can be written for $\epsilon > 0$,

$$\begin{aligned}\frac{1}{\sigma_n^2}\sum_{i=1}^n \mathbb{E}\left[Y_i^2 \mathbb{I}_{[Y_i^2 > \epsilon\sigma_n^2]}\right] &\le \frac{n}{\sigma_n^2}\mathbb{E}\left[Y_i^2 \mathbb{I}_{[Y_i^2 > \epsilon\sigma_n^2]}\right] \\ &\le \frac{C\cdot n}{\sigma_n^2}\mathbb{P}\left[\left[\mathbb{I}_{(X_i\le\xi)} - F(\xi)\right]^2 > \epsilon\sigma_n^2\right] \\ &\le \frac{C\cdot n}{\epsilon\sigma_n^4}\mathbb{E}\left[\left[\mathbb{I}_{(X_i\le\xi)} - F(\xi)\right]^2\right] \to_{n\to\infty} 0.\end{aligned}$$

The conditions:

(A) $\sigma_n^2 = nh(n)$ où $h(n)$ is a slowly varying function defined on $\mathbb{R}$,

(B) $\sup_{m\ge0,n\ge1}\left[\mathbb{E}\left(\sum_{i=1}^{m+n} Y_i - \sum_{i=1}^m Y_i\right)^2/\sigma_n^2\right] < \infty$,

therein are a result of $0 < \sigma_\infty^2 < \infty$. We take $t = 1$ to conclude

$$\sqrt{n}\left(F_n(\xi) - F(\xi)\right) \to^d \mathcal{N}\left(0, \sigma_\infty^2(\xi)\right).$$

Therefore by Theorem 2

$$\sqrt{n}\left(\xi_n - \xi\right) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma_\infty^2(\xi)}{f^2(\xi)}\right).$$

4.2 $\rho-$mixing Process

For a stationary sequence $(X_i)_{i\in\mathbb{Z}}$,we define

$$\begin{aligned}\alpha_n^* &= \sup_{S,T\subset\mathbb{Z},dist(S,T)\ge n} \alpha\left(\mathcal{M}_T, \mathcal{M}_S\right), \\ \rho_n^* &= \sup_{S,T\subset\mathbb{Z},dist(S,T)\ge n} \rho\left(\mathcal{M}_T, \mathcal{M}_S\right),\end{aligned}$$

where $\mathcal{M}_T = \sigma(X_i, i\in T), \mathcal{M}_S = \sigma(X_i, i \in S)$. We apply a result by Utev and Peligrad (2003, Theorems 2.1 and 2.2) to prove the following Theorems:

Corollary 2 *Under condition $H(X)$, we assume: $H(\rho)$: There exists a real number $0 \le \eta < 1$ and integer number $N \ge 1$ such that $\rho_N^* \le \eta$. Then, for $p > 2$ and $\delta > n^{-\frac{1}{2}\left(1-\frac{1}{p}\right)}$, there is a positive constant $C(p,N,\eta)$ such that for any $f \in \mathcal{F}$*

$$\mathbb{E}\left[\sup_{\|f_x - f_y\|_2\le\delta}\left|Z_n\left(f_x - f_y\right)\right|\right] \le C(p,N,\eta).\left(\ln n\cdot n^{\frac{1}{p}-\frac{1}{2}} + \delta^{\left(1-\frac{2}{p}\right)}\right).$$

If $H(F)$ is verified, then for $a_n \gg n^{-\frac{1}{2}}$ we have

$$|\xi_n - \xi| = o_p(a_n) \qquad \textit{in probability.}$$

If in addition the sequence $(X_i)_{i\ge1}$ is stongly mixing and $0 < \sigma_\infty^2 < \infty$, then

$$\sqrt{n}\left(\xi_n - \xi\right) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma_\infty^2(\xi)}{f^2(\xi)}\right).$$

Proof of Corollary 2. Assuming that the condition $H(\rho)$ is satisfied and the random variables are identically distributed, then by Utev and Peligrad (2003, Theorem 2.1), for any $p > 2$, there exists a positive constant $D = D(p,N,\eta)$ such that for $n \ge 1$,

$$\mathbb{E}\left|Z_n(f)\right|^p \le D\left(\|f\|_2^p + n^{1-\frac{p}{2}}\|f\|_p^p\right).$$

Apply now Theorem 1 with the condition $H(p,X)$ where $\varepsilon = 0$, $v = 2$ and $p = r$, we obtain

$$\mathbb{E}\left[\sup_{\|f_x - f_y\|_2\le\delta}\left|Z_n\left(f_x - f_y\right)\right|\right] \le C\cdot\left(\ln n\cdot n^{\frac{1}{p}-\frac{1}{2}} + \delta^{\left(1-\frac{2}{p}\right)}\right).$$

If $H(F)$ is verified and $a_n \gg n^{-\frac{1}{2}}$, then by Lemma 1 for $p > 2$ we obtain

$$|\xi_n - \xi| = o_p(a_n) \qquad \text{in probability.}$$

To show that

$$\sqrt{n}\left(F_n(\xi) - F(\xi)\right) \to^d \mathcal{N}\left(0, \sigma_\infty^2(\xi)\right),$$

we will apply a result by Utev and Peligrad (2003, Theorem 2.2, p. 105) with $\xi_{ni} \equiv \mathbb{I}_{(X_i \leq \xi)} - F(\xi)$, $\sigma_n^2 = \mathbb{E}\left[\sum_{i=1}^{n} \xi_{ni}\right]^2$, $k_n = n$ and $W_n(t) := \frac{1}{\sigma_n}\sum_{i=1}^{v_t} \xi_{ni}$ where $v_t = [nt]$ and $t \in [0,1]$. Si on a $0 < \sigma_\infty^2 < \infty$, alors $\frac{\sigma_n^2}{n} \to_{n\to\infty} \sigma_\infty^2$. The condition (2.5) of Utev and Peligrad (2003):

(2.5) $\lim_{n\to\infty} \sup\left[n\mathbb{E}(\xi_{n1})^2 / \sigma_n^2\right] \leq C$, is a consequence of $0 < \sigma_\infty^2 < \infty$. The condition (2.3) is proved in Corollary 2: (condition (L). We take $t = 1$ to conclude

$$\sqrt{n}\left(F_n(\xi) - F(\xi)\right) \to^d \mathcal{N}\left(0, \sigma_\infty^2(\xi)\right).$$

Therefore by Theorem 2

$$\sqrt{n}\left(\xi_n - \xi\right) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma_\infty^2(\xi)}{f^2(\xi)}\right).$$

4.3 $\alpha-$mixing Process

Corollary 3 *Under conditions $H(X)$ and $H(F)$, if the $\alpha-$mixing coefficient satisfies*

$$\alpha(n) \leq Cn^{-\theta} \text{ for some } C \geq 1 \text{ and } \theta > 1 + \sqrt{2}.$$

Then, for $a_n \gg n^{-\frac{1}{2}}$ we have

$$|\xi_n - \xi| = o_p(a_n).$$

and if in addition $0 < \sigma_\infty^2 < \infty$, then

$$\sqrt{n}\left(\xi_n - \xi\right) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma_\infty^2(\xi)}{f^2(\xi)}\right).$$

Proof of Corollary 3. When $(X_i)_{i\geq 1}$ are identically distributed, then by Shao and Yu (1996, Theorem 4.1), if

$$\alpha(n) \leq Cn^{-\theta} \text{ for } C > 0 \text{ and } \theta > 0.$$

Then, for some real numbers $2 < p < r \leq \infty, 2 < v \leq r, \varepsilon > 0$, $\theta > \frac{v}{v-2}$ and $\theta \geq \frac{(p-1)r}{r-p}$, there is a constant $K = K(v, p, r, \varepsilon, \theta, C) < \infty$ such that for any $f \in \mathcal{F}$

$$\mathbb{E}\left|Z_n(f)\right|^p \leq K\left(\|f\|_v^p + n^{1+\varepsilon-\frac{p}{2}}\|f\|_r^p\right)$$

which satisfies $H(p, X)$. If $\varepsilon \leq \frac{p}{2} - 1$ and $a_n \gg n^{-\frac{1}{2}}$, then by Lemma 1 we have

$$|\xi_n - \xi| = o_p(a_n).$$

For determining θ which allows to apply Theorem 1 we need $v < p < r$ and $\frac{p}{2}\left(1 - \frac{1}{p} + \frac{1}{r}\right) - 1 > 0$. Now we have

$$\theta \geq \frac{(p-1)r}{r-p} \Leftrightarrow p \leq \frac{r(\theta+1)}{\theta+r},$$

and

$$\theta > \frac{v}{v-2} \Leftrightarrow v > \frac{2\theta}{\theta-1}.$$

Since $v < p$ we need

$$\frac{2\theta}{\theta-1} < \frac{r(\theta+1)}{\theta+r}$$

which is satisfied if

$$\theta > 1 + \sqrt{2}\left(\frac{\sqrt{r(r-1)} + \sqrt{2}}{r-2}\right).$$

Consequently, we take $\theta = 1 + \eta$ where $\eta > \sqrt{2}$. For $\eta > \sqrt{2}$ we have $2 + \frac{2}{\eta} < 2 + \frac{2\eta^2+6\eta+4}{\eta^3+\eta^2+2\eta+4} < 2 + \frac{2(2\eta+3)}{\eta^2-2}$, then we choose v, p, r

i) $v = 2 + \frac{2}{\eta}$,

ii) $p = 2 + \frac{2\eta^2+6\eta+4}{\eta^3+\eta^2+2\eta+4}$,

iii) $r = 2 + \frac{2(2\eta+3)}{\eta^2-2}$.

With these choices we have $v < p < r$ and

$$\frac{p}{2}\left(1 - \frac{1}{p} + \frac{1}{r}\right) - 1 > 0.$$

Then we have

$$W(n, a_n) \leq C.\left(n^{-\frac{1}{2}+\frac{2+\varepsilon-\frac{p}{r}}{p+1-\frac{p}{r}}} + a_n^{\frac{1}{v}-\frac{1}{p}}\right) \rightarrow_{n\rightarrow\infty} 0.$$

If in addition $0 < \sigma_\infty^2 < \infty$, then by Rio (1997, Theorem 7.2) for

$$\alpha(n) \leq Cn^{-\theta} \text{ where } C \geq 1 \text{ and } \theta > 1,$$

we have

$$\sqrt{n}\left(F_n(\xi) - F(\xi)\right) \rightarrow^d \mathcal{N}\left(0, \sigma_\infty^2(\xi)\right).$$

Finally, by applying Theorem 2 for $a_n \gg n^{-\frac{1}{2}}$, we obtain that

$$\sqrt{n}\left(\xi_n - \xi\right) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma_\infty^2(\xi)}{f^2(\xi)}\right).$$

4.4 Nonlinear Functional of Gaussian Sequences

Corollary 4 *Let $X_i = G(Z_i)$ where G is a measurable function and (Z_i) is a stationary Gaussian sequence with zero mean and covariance function*

$$\varrho(n) = E(Z_i Z_{i+n}).$$

Assume $\sum_{i=0}^{\infty} |\varrho(i)| < \infty$. Then, for $p > 2$ and $\delta > n^{-\frac{1}{2}\left(1-\frac{1}{p}\right)}$, there is a positive constant $C(p, \varrho)$ such that for any $f \in \mathcal{F}$

$$\mathbb{E}\left[\sup_{\|f_x - f_y\|_2 \leq \delta} \left|Z_n\left(f_x - f_y\right)\right|\right] \leq C(p, \varrho).\left(\ln n.n^{\frac{1}{p}-\frac{1}{2}} + \delta^{\left(1-\frac{2}{p}\right)}\right).$$

If $H(F)$ is verified, then for $a_n \gg n^{-\frac{1}{2}}$ we have

$$|\xi_n - \xi| = o_p(a_n) \qquad \textit{in probability,}$$

and if in addition $0 < \sigma_\infty^2 < \infty$, then

$$\sqrt{n}\left(\xi_n - \xi\right) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma_\infty^2(\xi)}{f^2(\xi)}\right).$$

Proof of Corollary 4. The proof of this corollary is a consequence of the following results:

Lemma 3 (Ben Hariz, 2011) *Let p be an even integer and assume that $\sum_{i=0}^{\infty} |\varrho(i)| < \infty$, then there exists a constant $K = K(p, \varrho)$ such that for all $n > 0$,*

$$\mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f(Z_i) - E(f(Z_i))\right)^p \leq K(p, \varrho)\left(\|f\|_2^p + n^{1-\frac{p}{2}}\|f\|_p^p\right).$$

We apply Lemma 3 for $f(Z) = \mathbb{I}_{G(Z)\leq x}$. Then $H(p, X)$ is satisfied with $\varepsilon = 0, v = 2$ and $p = r$. If $H(F)$ is verified, then by Lemma 1 for $a_n \gg n^{-\frac{1}{2}}$, we have

$$|\xi_n - \xi| = o_p(a_n).$$

And by Theorem 1

$$W(n,\delta) \leq C.\left(\ln n.n^{-\frac{1}{2}+\frac{1}{p}} + \delta^{\left(1-\frac{2}{p}\right)}\right).$$

For the central limit theorem we need to apply the following results due to Breuer and Major (1983), (see also Csörgo, Sándor & Mielniczuk, 1996, for a functional extension) .

Lemma 4 *Let* (Z_i) *be a stationary Gaussian sequence with a covariance function satisfying* $\sum_{i=0}^{\infty} |\varrho(i)| < \infty$, *then*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbb{I}_{G(Z_i)\leq x} - F(x)\right) \longrightarrow^{d} \mathcal{N}\left(0,\sigma_{\infty}^{2}(x)\right)$$

where $\sigma_n^2(x) = Var\left(\mathbb{I}_{G(Z_i)\leq x}\right) + 2\sum_{i=1}^{\infty} Cov\left(\mathbb{I}_{G(Z_1)\leq x}, \mathbb{I}_{G(Z_i)\leq x}\right)$.

If $0 < \sigma_{\infty}^{2} < \infty$, then by Lemma 3, Lemma 4 and Theorem 2 we have

$$\sqrt{n}\left(\xi_n - \xi\right) \longrightarrow^{d} \mathcal{N}\left(0, \frac{\sigma_{\infty}^{2}(\xi)}{f^{2}(\xi)}\right).$$

4.5 Comparison with the Existing Results of the Literature

- In Sen (1972), Sen has proved that for a φ-mixing sequence of random variables, if we have

$$\sum_{i=0}^{\infty}\varphi^{\frac{1}{2}}(i) < \infty,$$

then

$$\sqrt{n}\left(\xi_n - \xi\right) \longrightarrow^{d} \mathcal{N}\left(0, \frac{\sigma_{\infty}^{2}(\xi)}{f^{2}(\xi)}\right).$$

which is stronger than our condition:

$$\sum_{i=0}^{\infty}\varphi^{\frac{1}{p}}\left(2^{i}\right) < \infty.$$

Indeed, $\sum_{i=0}^{\infty}\varphi^{\frac{1}{2}}(i) < \infty$ needs an algebraic decay of the the mixing coefficient $\varphi(i)$, and $\sum_{i=0}^{\infty}\varphi^{\frac{1}{p}}\left(2^{i}\right) < \infty$ needs only a logarithmic decay.

- In 2005, Chen and Tang studied the nonparametric estimation of the Value at Risk (VaR) for a geometric α-mixing sequence of random variables, that means

$$\alpha(k) \leq c\rho^{k} \text{ where } k \geq 1, c > 0 \text{ and } \rho \in (0,1).$$

Using the kernel estimation of the VaR:

$$\widehat{F}_{n,h}\left(\widehat{VaR}_h(q)\right) = \frac{1}{n}\sum_{i=1}^{n} G\left(\frac{\widehat{VaR}_h(q) - X_i}{h}\right) = q,$$

where $G(x) = \int_{-\infty}^{x} K(u)\,d(u)$ is a distribution function of a kernel density K, they showed that:

$$\left|\widehat{VaR}_h(q) - VaR(q)\right| = o_{a.s.}\left(n^{-\frac{1}{2}}\ln(n)\right).$$

$$\sqrt{n}\left(\widehat{VaR}_h(q) - VaR(q)\right) \rightarrow {}^{d}\mathcal{N}\left(0, \frac{\sigma_{\infty}^{2}(VaR(q))}{f^{2}(VaR(q))}\right).$$

- Lahiri and Sun (2009) showed that for a α-mixing sequence of random variables such that

$$\alpha(n) \leq dn^{-\theta} \text{ where } \theta > 12,$$

the empirical $\widehat{VaR}(q)$ satisfy for a constant $C > 0$ and $n \geq 1$

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left[\sqrt{n}\left(\xi_n - \xi\right) \leq x\right] - \Phi\left[x \times \frac{f(\xi)}{\sigma_{\infty}(\xi)}\right]\right| \leq \frac{C}{\sqrt{n}},$$

where Φ is the standard normal distribution. In particular they obtained as $n \to \infty$,

$$\sqrt{n}\,(\xi_n - \xi) \longrightarrow^d \mathcal{N}\left(0, \frac{\sigma_\infty^2(\xi)}{f^2(\xi)}\right).$$

Observe that for the CLT to hold for strong mixing sequences, we only need that $\alpha(k) \leq Cn^{-\theta}$ with $\theta > 1 + \sqrt{2}$.

Remark 2 Our results also apply for stochastic differential equations and stochastic volatility models discretely observed. Indeed, Genon-Catalot et al. (2000) showed that, under some conditions, these models as well as theirs discrete versions, satisfies geometric α or $\rho-$mixing. Therefore the main hypothesis $H(p, X)$ is then fulfilled for any $p \geq 2$. Regarding GARCH models which are also widely used in financial modeling, we mention that Davis et al. (1999) showed that under conditions on the moment of the innovations and on the Lyapunov exponent associated to the sequence, the squared of the GARCH sequence is geometric $\alpha-$mixing. Hence, our results apply also for GARCH models.

5. Simulation Studies

In this section we present some numerical studies which illustrate the conditions under which $\widehat{VaR}(q)$ converges to $VaR(q)$. In these simulations, we choose a correlated Gaussian and Pareto sequences. In both cases, we compare the $VaR(q)$ where $q = 0.95$ to the empirical estimate of $VaR(q)$. For each set of parameters, we run ($M = 10000$) Monte Carlo simulations and compute the mean absolute error ($MAE(n)$) between $\widehat{VaR}(q)$ and $VaR(q)$

$$MAE(n) = \frac{1}{M}\sum_{i=1}^{M}\left|\widehat{VaR}_{(i)}(q) - VaR(q)\right|.$$

We also give a confidence interval with level 95% to the $VaR(q)$. We consider three different models. First, a correlated Gaussian sequence, then a correlated sequences with Pareto marginal distributions and finally a stochastic volatility model.

5.1 Case 1: Dependent Gaussian Process

Let $(X_i)_{0\leq i\leq n}$ be a Gaussian sequence with zero mean, unit variance and a correlation function given by:

$$\varrho_n(i) := Cov(X_0, X_i) = (1 + |i|)^{-\alpha} \quad, i = 1, ..., n$$

where $\alpha > 0$. The parameter α tunes the strength of dependence. In particular $\alpha = \infty$ corresponds to the i.i.d. sequence, whereas $\alpha = 0, (\varrho_n(i) = 1)$ gives perfectly correlated sequence.

We study the process:

$$T_n := \sqrt{n}\left(\widehat{VaR}(q) - VaR(q)\right).$$

We show that for $\alpha > 1 \left(\Rightarrow \sum_{i=0}^{\infty} |\varrho_n(i)| < \infty.\right)$,

$$T_n \longrightarrow^d_{n\to\infty} \mathcal{N}\left(0, \tau_\infty^2\right), \tag{5.1}$$

where $\tau_\infty^2 = \frac{\sigma_\infty^2(VaR(q))}{f^2(VaR(q))}$. Here we recall that $VaR(0.95) = 1.6449$.

In Figure 1, we plot the mean absolute error with a 95% confidence interval as a function of n for different values of α when $q = 0.95$. Clearly the $MAE(n)$ goes to zero when n large, for any $\alpha > 0$. The simulations shows that the $\widehat{VaR}(q)$ is consistent when the correlation parameter $\alpha > 0$. When $\alpha > 1$, in Figure 2, we plot $\sqrt{n}\,MAE$ against n to see that it converges to a constant. In Figure 3, we see that the $MAE(n)$ as a function of α for different values of n with $q = 0.95$, tends to zero for large values of n. In Figure 4, we compare the histogram of T_n for $\alpha = 3$ and $n = 800$ with the density function of Gaussian distribution $\mathcal{N}\left(0, \tau_\infty^2\right)$. Clearly, for $\alpha > 1$ the histogram of T_n is close to the normal distribution, confirming our result (5.1).

5.2 Case 2: Dependent Pareto Process

We now consider the $\widehat{VaR}(q)$ for a correlated Pareto sequence $(X_i)_{0\leq i\leq n}$. Recall that the distribution function of Pareto is defined for $\beta > 0$ by:

$$G_\beta(x) = \begin{cases} 1 - \left(\frac{x_0}{x}\right)^\beta & x > x_0 \\ 0 & x \leq x_0 \end{cases}$$

To construct a correlated Pareto sequence, we let $X_i = G_\beta^{-1}(\Phi(Y_i))$ where Φ is the Gaussian distribution $\mathcal{N}(0,1)$ and $\{Y_i\}_{0\leq i\leq n}$ is a correlated Gaussian sequence defined as in the previous example. As in the first case, we study the process T_n to illustrate the central limit theorem (see (5.1)). Here $VaR(0.95) = 2.7144$ when $\beta = 3$.

In Figure 5, we plot $MAE(n)$ with a 95% confidence interval as a function of n for different values of α when $q = 0.95$. Clearly, the MAE goes to zero when n large, for any $\alpha > 0$. The simulations shows that the $\widehat{VaR}(q)$ is consistent when the correlation parameter $\alpha > 0$. When $\alpha > 1$, in Figure 6, we plot $\sqrt{n}\ MAE(n)$ against n to see that it converges to a constant. In Figure 7, we see that the $MAE(n)$ as a function of α for different values of n with $q = 0.95$, tends to zero for large values of n. In Figure 8, we compare the histogram of T_n for $\alpha = 3$ and $n = 800$, with the density function of Pareto distribution. Here again, when $\alpha > 1$, the CLT is satisfied.

5.3 Case 3: Stochastic Volatility Models

We assume that $\widehat{VaR}(q)$ of the correlated sequence $(X_i)_{0\leq i\leq n}$ with stochastic volatility:

$$X_i = \sigma_i.\varepsilon_i$$

where $(\varepsilon_i)_{0\leq i\leq n}$ is an iid Gaussian sequence $\mathcal{N}(0,1)$ and $(\sigma_i)_{0\leq i\leq n}$ correlated Gaussian or Pareto sequences.

As in the first case, we study the process T_n to prove (5.1) where $VaR(0.95) \approx 1.5949$ for the Gaussian sequence and $VaR(0.95) \approx 2.4615$ for the Pareto sequence with $\beta = 3$. In Figure 9, we compare the histogram of T_n for $\alpha = 3$ and $n = 800$, with the density function of Gaussian distribution $\mathcal{N}\left(0, \tau_\infty^2\right)$ using two cases (Gaussian and Pareto for the distribution function of σ_i). Here again, when $\alpha > 1$, the CLT is satisfied.

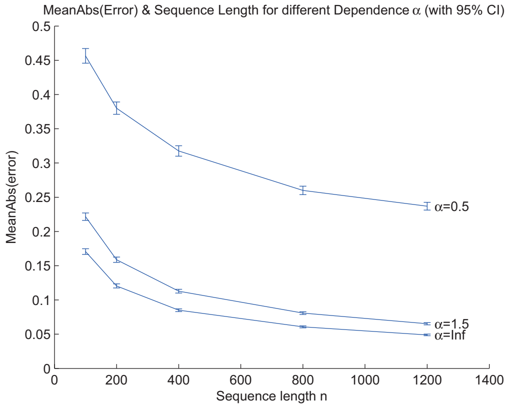


Figure 1. The Mean Absolute Error ($MAE(n)$) with 95% confidence intervals for correlated Gaussian sequence with correlation function $\varrho_n(i) = (1+|i|)^{-\alpha}$ is plotted against the sequence length n for different values of dependence parameter α

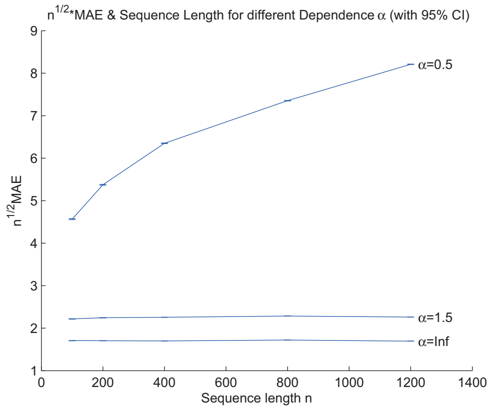


Figure 2. $\left(\sqrt{n}MAE(n)\right)$ with 95% confidence intervals for correlated Gaussian sequence with correlation function $\varrho_n(i)$ is plotted against the sequence length n for $\alpha \in \{0.5, 1.5, \infty\}$. The value $\left(\sqrt{n}MAE(n)\right)$ tends to a constant for $\alpha > 1$ indicating that the optimal convergence rate $O(n^{-\frac{1}{2}})$ is achieved

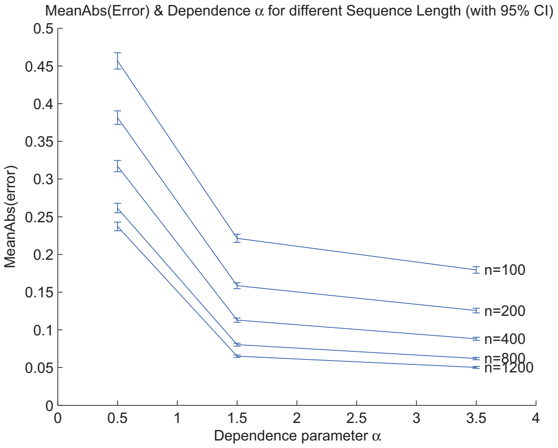
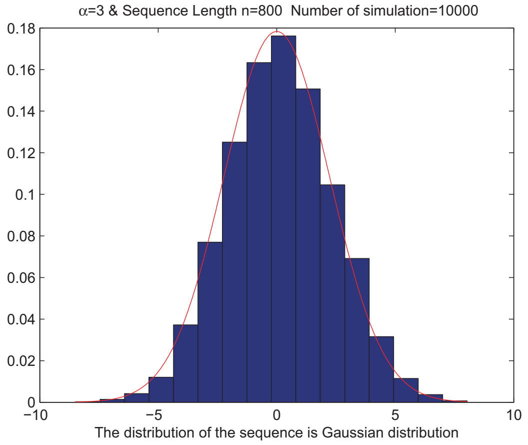


Figure 3. The Mean Absolute Error ($MAE(n)$) with 95% confidence intervals for correlated Gaussian sequence with correlation function $\varrho_n(i)$ is plotted against the dependence parameter α for different values of n



Figue 4. Comparing the histogram of T_n for a Gaussian sequence where $\alpha = 3$ and $n = 800$, with the density function of Gaussian distribution $\mathcal{N}\left(0, \tau_\infty^2\right)$

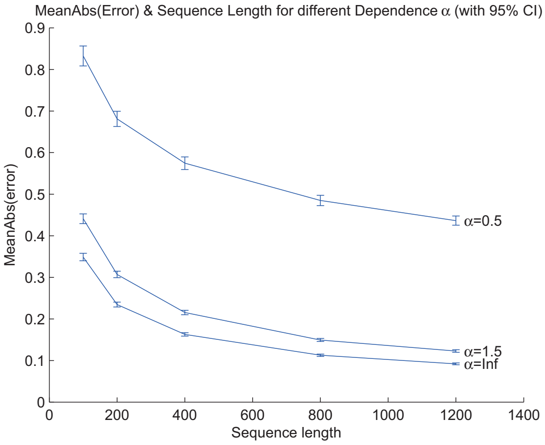


Figure 5. The Mean Absolute Error ($MAE(n)$) with 95% confidence intervals for correlated **Pareto** sequence with correlation function $\varrho_n(i)$ is plotted against the sequence length n for different values of dependence parameter α

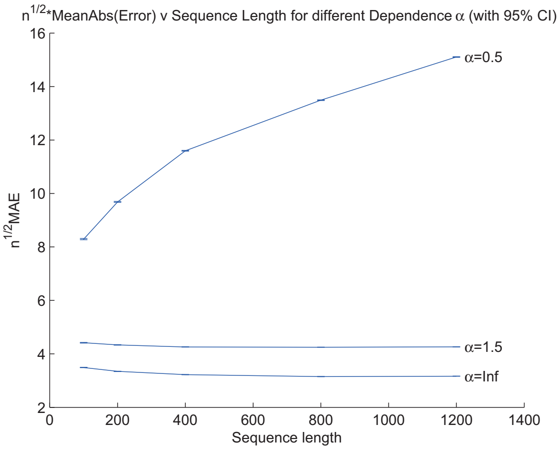


Figure 6. $\left(\sqrt{n}MAE(n)\right)$ with 95% confidence intervals for correlated **Pareto** sequence with correlation function $\varrho_n(i)$ is plotted against the sequence length n for $\alpha \in \{0.5, 1.5, \infty\}$. The value $\left(\sqrt{n}MAE(n)\right)$ tends to a constant for $\alpha > 1$ indicating that the optimal convergence rate $O(n^{-\frac{1}{2}})$ is achieved

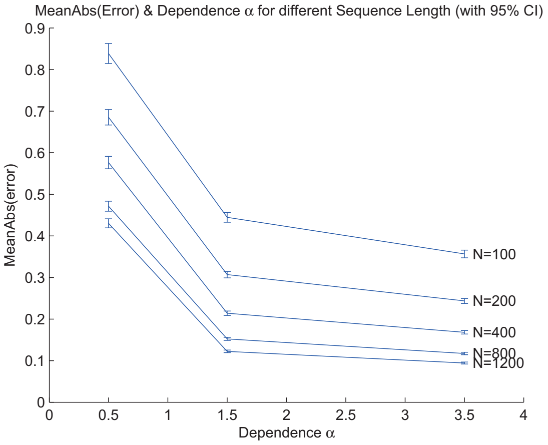


Figure 7. The Mean Absolute Error ($MAE(n)$) with 95% confidence intervals for correlated **Pareto** sequence with correlation function $\varrho_n(i)$ is plotted against the dependence parameter α for different values of n

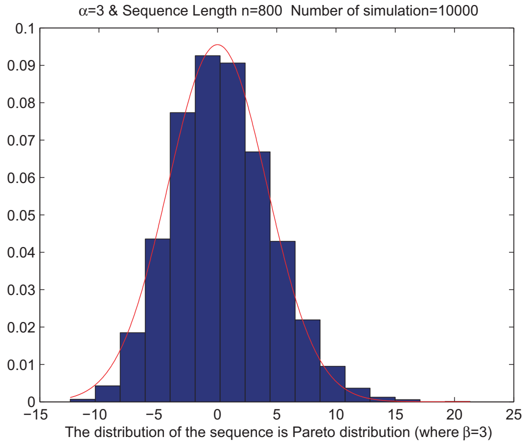


Figure 8. Comparing the histogram of T_n for a Pareto sequence where $\alpha = 3$ and $n = 800$, with the density function of Gaussian distribution $\mathcal{N}\left(0, \tau_\infty^2\right)$

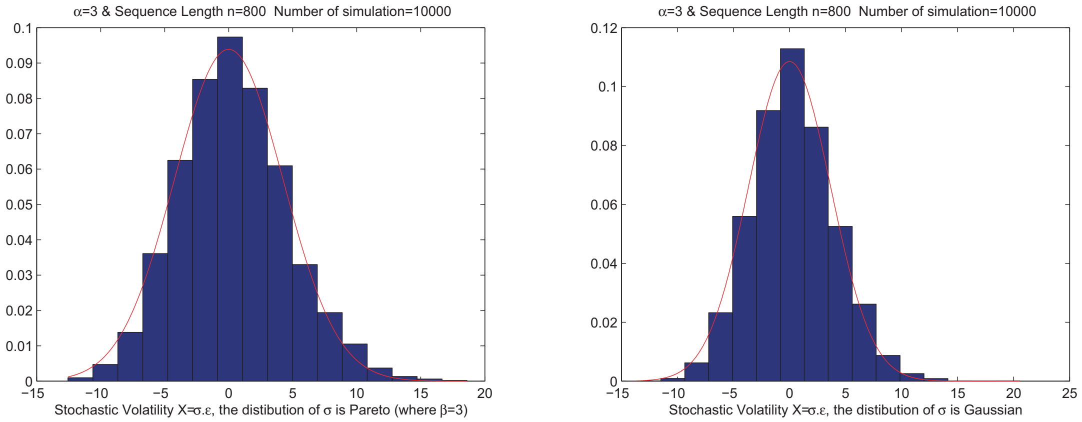


Figure 9. Comparing the histogram of T_n for $\alpha = 3$ and $n = 800$, with the density function of Gaussian distribution $\mathcal{N}\left(0, \tau_\infty^2\right)$ for two case (Gaussian and Pareto sequence)

6. Conclusion

In this work, we considered the nonparametric estimator of the VaR. We proved the consistency of the empirical estimator and a central limit theorem for $\sqrt{n}\,(\xi_n - \xi)$. Ours results apply as soon as we have a moment inequality for the partial sums. Although the limit is normal like the i.i.d. case, the limiting variance is different and typically larger with dependent observations. One consequence is: the confidence interval for the VaR will be larger. Another question arise about the estimation of this variance. Our results apply for weakly dependent sequences, including mixing sequences, linear process, gaussian sequences and others. It would be interesting to study the estimation of the VaR for long-range dependent sequences.

References

Ben Hariz, S. (2005). Uniform CLT for empirical process. *Stochastic Processes and their Applications, 115*(2), 339-358. http://dx.doi.org/10.1016/j.spa.2004.09.006

Ben Hariz, S. (2011). *Moment Inequalities for short and long-range dependent sequences.* Preprint.

Bennett, G. (1962). Probability Inequalities for the Sum of Independent Random Variables. *Journal of the American Statistical Association 57*(297), 33-45. http://dx.doi.org/10.2307/2282438

Breuer, P., & Major, P. (1983). Central limit theorems for nonlinear functionals of Gaussian fields. *J. Multivariate Anal., 13*(3), 425-441. http://dx.doi.org/10.1016/0047-259X(83)90019-2

Chen, S. X., & Tang, C. Y. (2005). Nonparametric Inference of Value at Risk for dependent Financial Returns. *Journal of Financial Econometrics, 3*, 227-255.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance, 1*, 223-236. http://dx.doi.org/10.1080/713665670

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, New Jersey: Princeton University Press.

Csörgo, S., & Mielniczuk, J. (1996). The empirical process of a short-range dependent stationary sequence under Gaussian subordination. *Probab. Theory Related Fields, 104*(1), 15-25. http://dx.doi.org/10.1007/BF01303800

Davis, R. A., Mikosch, T., & Basrak, B. (1999). *Sample ACF of multivariate stochastic recurrence equations with applications to GARCH.* Preprint.

Dowd, K. (2001). Estimating VaR with order statistics. *Journal of Derivatives*, 23-30. http://dx.doi.org/10.3905/jod.2001.319154

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association, 58*(301), pp. 13-30.

Hu, S. H. (2003). Some new results for the strong law of large numbers. *Acta Mathematica Sinica, 46*, 1123-1134.

Lahiri, S. N., & Sun, S. (2009) A Berry-Esseen theorem for sample quantiles under weak dependence. *Ann. Appl.*

Probab., 19(1), 108-126. http://dx.doi.org/10.1214/08-AAP533

Marinelli, C., d'Addona, S., & Rachev, S. T. (2007). A comparison of some univariate models for value-at-risk and expected shortfall. *Int. J. Theor. Appl. Finance, 10*(6), 1043-1075. http://dx.doi.org/10.1142/S0219024907004548

Peligrad, M. (1985). An invariance principle for φ-mixing sequences. *Ann. Probab. 13*(4), 1304-1313. http://dx.doi.org/10.1214/aop/1176992814

Rio, E. (1997). Théorèmes limites pour des variables aléatoires faiblement dépendentes. Preprint, Universit de Paris Sud, no. 97/81.

Sen, P. K. (1972). On the Bahadur representation of sample quantiles for sequences of ϕ-mixing random variables. *J. Multivariate Anal., 2*, 77-95.

Shao, Q., & Yu, H. (1996). Weak convergence for weighted empirical processes of dependent sequences. *Ann. Probab., 24*(4), 2098-2127. http://dx.doi.org/10.1214/aop/1041903220

Utev, S., & Peligrad, M. (2003). Maximal inequalities and an invariance principle for a class of weakly dependent random variables. *J. Theoret. Probab., 16*(1), 101-115.

Valentine Genon-Catalot, Thierry Jeantheau, & Catherine Larédo. (2000). Stochastic Volatility Models as Hidden Markov Models and Statistical Applications. *Bernoulli,* 6(6), 1051-1079. http://dx.doi.org/10.2307/3318471

Van der Vaart, Aad W., & Wellner, Jon A. (1996). *Weak convergence and empirical processes. With applications to statistics*. Springer Series in Statistics. New York: Springer-Verlag. ISBN: 0-387-94640-3.

International Journal of Statistics and Probability; Vol. 2, No. 1; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Analysis of Multiple Myeloma Life Expectancy Using Copula

Seung-Hwan Lee[1], Phillip Yulin Deng[1] & Eun-Joo Lee[2]

[1] Department of Mathematics, Illinois Wesleyan University, Bloomington, USA

[2] Department of Mathematics, Millikin University, Decatur, USA

Correspondence: Seung-Hwan Lee, Department of Mathematics, Illinois Wesleyan University, Bloomington, IL 61702, USA. Tel: 1-309-556-3421. E-mail: slee2@iwu.edu

Received: September 10, 2012 Accepted: January 7, 2013 Online Published: January 16, 2013

doi:10.5539/ijsp.v2n1p44 URL: http://dx.doi.org/10.5539/ijsp.v2n1p44

Abstract

Multiple myeloma is a blood cancer that develops in the bone marrow. It is assumed that in most cases multiple myeloma develops in association with several medical factors acting together, although the leading cause of the disease has not yet been identified. In this paper, we investigate the relationship between the factors to measure multiple myeloma patients' survival time. For this, we employ a copula that provides a convenient way to construct statistical models for multivariate dependence. Through an approach via copulas, we find the most influential medical factors that affect the survival time. Some goodness-of-fit tests are also performed to check the adequacy of the copula chosen for the best combination of the survival time and the medical factors. Using the Monte Carlo simulation technique with the copula, we re-sample survival times from which the anticipated life span of a patient with the disease is calculated.

Keywords: elliptical copula, multiple myeloma, tail dependence

1. Introduction

Multiple myeloma is a blood cancer caused by the accumulation of abnormal plasma cells in the bone marrow. It is the second most common blood cancer, after non-Hodgkin's lymphoma, and represents approximately 1 percent of all cancers and 2 percent of all cancer deaths. The American Cancer Society estimates that 20,180 people were diagnosed with multiple myeloma during 2010 (Multiple Myeloma: Disease Overview, Multiple Myeloma Research Foundation, 2010). The prognosis of the disease is often unpredictable and overall survival is ranged from a few months to more than 10 years (Kyle & Rajkumar, 2008). It has been known that multiple myeloma may be the result of several medical risk factors that act together. Therefore, it is important to understand the relationship between survival time of patients and the risk factors in the bone marrow for the disease. So this work investigates possible associations in multiple myeloma data, carried out at the Medical Centre at the University of West Virginia, with a particular focus on potential survival time (denoted by ST hereafter) of a patient over the treatment period in association with the following four medical factors: level of blood urea nitrogen (BUN), serum calcium (CA), hemoglobin (HB), and the percentage of plasma cells (PC). See Collect (1999) and Krall et al. (1975) for the details of the data.

When analyzing dependence of several variables, such as the survival time and associated disease factors, an often used measure is Pearson's linear correlation coefficient. However, the linear correlation coefficient does not detect non-linear behaviors of the variables considered, is strongly affected by extreme values and not invariant under non-linear transformations (Embrechts, McNeil, & Straumann, 2002; Frees & Valdez, 1998; Schweizer & Wolff, 1981). To overcome these problems, we employ a statistical function called a copula which links multivariate distributions to marginal distributions. A copula captures both the linear and non-linear dependence that may exist between the variables. It is also invariant under monotone transformations. Especially, the use of a copula has gained importance as a simple tool to measure the amount of dependence in the tails. For example, the treatment of a disease can either exceed or fall below a given level at either the early or late stage of the disease. The use of a copula has been studied in several disciplines such as survival analysis (Zheng & Klein, 1995), risk management and financial applications (Breymann et al., 2003; Embrechts et al., 2002; 2003). Excellent reviews on copulas can be found in Nelson (1999).

Copulas have varying amounts of tail dependence. So an appropriate copula that fits the data should be used. A poorly chosen copula may lead to undesirable results. This issue has been studied by many authors, including

Melchiori (2003), Durrelaman (2000), Kumar and Shoukri (2008), Frees and Valdez (1998), and Genest and Rivest (1993). Similar to the approach employed in the literature, based on the distance of copula and its empirical version, we choose a copula that best fits data. We then assess the adequacy of the copula chosen. The procedures are based on the Monte Carlo simulation technique by which the distribution of the distance can be seamlessly approximated under the null hypothesis of the no model misspecification. As a numerical measure of the model adequacy, the empirical - value, obtained from the simulated distance process, is used.

The primary objective of this paper is to estimate the expected life span of a patient with the disease by identifying the dependence between the variables considered. To this end, a systematic study in search for the most influential medical factors that affects the survival time is performed through copulas. Based on a large number of simulated data obtained from a copula chosen for a best combination of the survival time and medical factors, we calculate maximum extension possible for a life with reference to the factors that influence the survival time. This work is organized as follows. An overview of copulas is given in section 2. Section 3 discusses the procedures of finding an appropriate copula for multiple myeloma data. Numerical results of the anticipated life span of a patient with the disease are presented in section 4. Section 5 concludes the paper.

2. Copula

If $X_1, \ldots, X_n$ are random variables, the multivariate distribution function is defined as

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n),$$

which completely describes the dependence between the random variables. A copula provides a useful way to construct such a multivariate distribution of two or more random variables. The essential idea of the copula approach is that a joint distribution is factored into two components: the marginal distributions and a dependence function called a copula, as described in the following theorem (Sklar, 1959).

Theorem *If F is a multivariate distribution function with marginal distribution functions $F_1, \ldots, F_n$, then there is a copula C such that*

$$F(x_1, \ldots, x_n) = C(F_1(x_1)), \ldots, F_n(x_n)). \tag{1}$$

Theorem states that a copula function defines a joint distribution, evaluated at $x_1, \ldots, x_n$, with marginal distributions $F_1, \ldots, F_n$. Letting f be the probability density function of F, $u_i = F_i(x_i)$, $i = 1, \ldots, n$, and c the density function of a copula C, it can be easily shown that

$$f(x_1, \ldots, x_n) = c(u_1, \ldots, u_n) \times \prod_i f_i(x_i).$$

This indicates that a multivariate probability density function $f(x_1, \ldots, x_n)$ can be split into the univariate marginal probability density functions $f_i(x_i)$'s and the copula density function $c(u_1, \ldots, u_n)$ that determines a dependence structure. Hence, it is possible to separately specify marginal density functions and the dependence relation determined by the copula density function. A copula itself is in fact a multivariate distribution with standard uniform marginal distributions due to the fact that $F_i(X_i)$'s are uniformly distributed over the interval [0,1]. So it maps points on the $n-$dimensional unit square to values between 0 and 1. Specifically, with the values $u_1, \ldots, u_n$ of standard uniform random variables $U_1, \ldots, U_n$, the copula $C(u_1, \ldots, u_n)$ of the multivariate distribution function F is

$$C(u_1, \ldots, u_n) = F(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)), \tag{2}$$

where $F_1^{-1}, \ldots, F_n^{-1}$ denote the quantile functions of the marginal distributions $F_1, \ldots, F_n$. Note that from (1) and (2), a specified copula determines the dependence structure of data, while the marginal distribution does not. This is due to the fact that a copula links univariate marginal distributions to their multivariate distribution and is independent of marginal distributions. Therefore, we need to consider several copulas, leading to different dependence structures, from which an appropriate copula for data is chosen. In this work, we look at elliptical copulas the most commonly used copulas.

Elliptical copulas are the copulas of elliptical distributions. Examples include the Gaussian and t copulas (Embrecht et al., 2002; Demnarta & McNeil, 2005). The key advantages of elliptical copulas are that they are suitable in modeling dependence structures with multi-dimensions and specify different levels of correlations between the marginal distribution functions. As seen from the expressions in (1) and (2), any marginal distributions can be imposed over an elliptical copula, provided that marginal distributions are known and can be consistently estimated from data (Nelson, 1999; Embrechts et al., 2002). A copula with given marginals is called meta-elliptical copula. However, for simplicity and without loss of generality, elliptical copula indicates meta-elliptical copula hereafter.

2.1 Gaussian and t Copulas

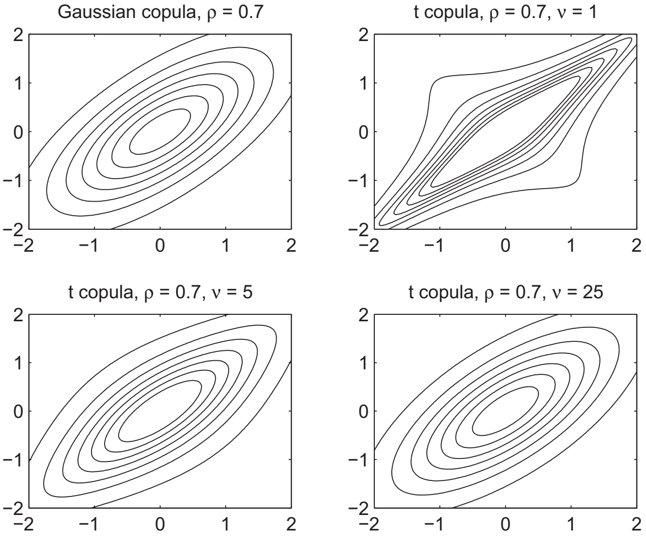


Figure 1. Contour plots of Gaussian and t copulas, $\rho = 0.7$

The Gaussian copula is the copula of multivariate normal distribution. From (2), it is defined as

$$C_G(u_1, \ldots, u_n) = \Phi_\rho(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)),$$

where Φ_ρ denotes the standard multivariate normal distribution function with correlation coefficient matrix ρ and Φ^{-1} the inverse of the standard univariate normal distribution function. As ρ approaches -1 and 1, the Gaussian copula captures stronger positive and negative linear relationship between random variables, respectively. See Figure 1 for these phenomena. The linear correlation coefficients depend on the marginal distributions, measuring the overall strength of a linear relationship, but give no information about how that varies across the distribution. So it is not preserved by copulas. It is important to note that the linear correlation coefficient has such shortcomings, but it is still needed to parameterize the Gaussian copula. Rank based correlations describe the global association of variables and are invariant under any monotonic transformations. In this work, the linear correlation coefficient (ρ) calculated by the Kendall's rank correlation (τ) was used such that $\rho = \sin(\tau\pi/2)$. We plug the marginal distributions into a Gaussian copula function to obtain a multivariate distribution. For example, using the copula $C_G(u_1, \ldots, u_n)$ and the marginal distribution functions $F_1, \ldots, F_n$, we have the multivariate distribution

$$F(x_1, \ldots, x_n) = C_G(F_1(x_1), \ldots, F_n(x_n)).$$

Note that the multivariate normal distribution has thin tails. Thus, the Gaussian copula is not appropriate in the analysis of tail dependence of variables that have heavy-tailed distributions (Embrechts et al., 2002). To repair this problem, we consider the t copula which can capture dependence between variables in the tails of the distribution. The t copula is a copula of the multivariate t distribution. From (2) it is defined as

$$C_t(u_1, \ldots, u_n) = t_{\rho,\nu}(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_n)),$$

where $t_{\rho,\nu}$ denotes the standard multivariate t distribution function with correlation coefficient matrix ρ and the degrees of freedom ν, and t_ν^{-1} is the inverse of the standard univariate t distribution function. Similar to the Gaussian copula, ρ is approximated by Kendall's τ. The second parameter ν controls the thickness of the tails of the distribution, exhibiting co-movement of the variables in the tails as seen in Figure 1.

Note that the Gaussian copula is a limiting case of the t copula as $\nu \to \infty$, and the t copula with $\nu = 1$ is often called the Cauchy copula. This implies that with increasing degrees of freedom, the t copula tends to look more like the Gaussian copula. We insert the marginal distributions into the t copula function to obtain a multivariate distribution, as follows.

$$F(x_1, \dots, x_n) = C_t(F_1(x_1), \dots, F_n(x_n)).$$

Table 1. Coefficient of Tail Dependence of the *t* copula (Embrechts et al., 2002)

$\nu \setminus \rho$	-0.5	0	0.5	0.9	1
2	0.06	0.18	0.39	0.72	1
4	0.01	0.08	0.25	0.63	1
10	0	0.01	0.08	0.46	1
∞	0	0	0	0	1

2.2 Tail Dependence

The relationship between random variables, especially in the tails, is controlled by the choice of copulas. It can be described by the coefficient of asymptotic tail dependence of a copula. For a pair of random variables, X and Y, it quantifies the probability to observe a large Y given that X is large. Specifically, the coefficient of upper tail dependence is

$$\lambda_U = \lim_{q \to 1-} P(Y > F_2^{-1}(q) | X > F_1^{-1}(q)),$$

provided the limit exists, and the coefficient of lower tail dependence is

$$\lambda_L = \lim_{q \to 0+} P(Y \leq F_2^{-1}(q) | X \leq F_1^{-1}(q)),$$

provided the limit exists. See Embrechts et al. (2002) for details on these expressions. Note that $\lambda_U = \lambda_L$ for the Gaussian and t copulas, due to the symmetric property of the elliptical distributions. Let $\lambda_U = \lambda_L = \lambda$. Then the following formula is often used for computational purposes:

$$\lambda = 2t_{\nu+1}(-\sqrt{\nu+1}\sqrt{1-\rho}/\sqrt{1+\rho}), \tag{3}$$

where t_ν is the standard univariate t distribution with ν degrees of freedom. Table 1 shows the results of the tail dependence coefficients for a few representative values of ρ and ν (Embrechts el al., 2002) calculated by the formula in (3).

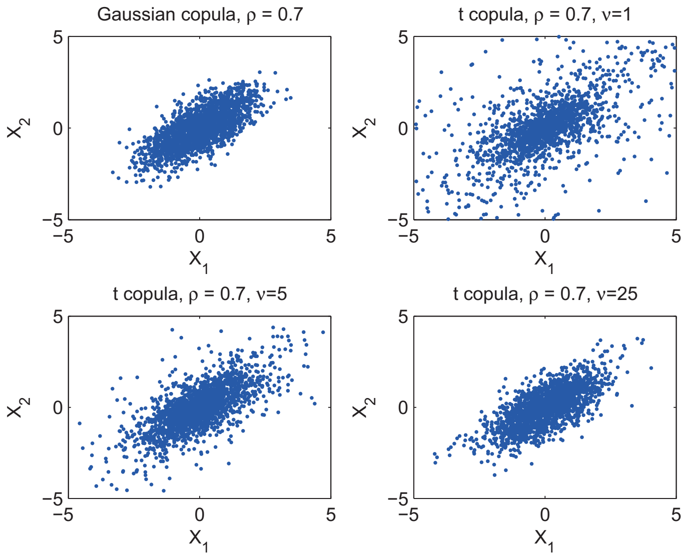


Figure 2. Scatter plots of 2000 simulated data points for Gaussian and *t* copulas

The tail dependence coefficient λ tends to zero as the degrees of freedom ν tends to infinity for $\rho < 1$. Since the Gaussian copula is a limiting case of the t copula as $\nu \to \infty$, the value of λ for the Gaussian copula is 0. So the

Gaussian copula exhibits no tail dependence. On the contrary, the t copula has non-negative values of λ for all values of ρ, and so the association of extreme values is captured by the t copula, with different amounts depending on ν at a fixed value of ρ. As seen in Table 1, λ increases as ν decreases at a fixed value of ρ. This indicates that the *t* copula has tail dependence that increases, whether it is upper tail dependence or lower tail dependence, with decreasing parameter ν. So it is useful when dependence of extreme values is observed in data.

Note that although the t copulas generate different dependence structures, they may still have the same marginal distribution functions and the same correlation. To illustrate this, we generate 2000 simulated data points using the Gaussian and t copulas with 1, 5, 25 degrees of freedom for ρ=0.7. The scatter plots of those simulated values are presented in Figure 2, displaying that the tail dependence is not remarkable in the t copula with ν=25 when compared to either ν=1 or ν=5. It is evident from this figure that the tail dependence becomes stronger as the degrees of freedom decreases, and this indicates that an increase in ν results in fewer occurrences of joint extremes. The plots also demonstrate that the dependence structure of multivariate distributions may not be perfectly identified only by their marginal distributions and correlations. In summary, a copula fully captures real dependence structure among random variables and hence provides a model that reflects on more detailed information about data.

3. Procedures

Multiple myeloma is a malignant disease caused by the accumulation of abnormal plasma cells in the bone marrow. Pain emanating from bone tissue and cancerous destruction of bone tissue may occur as a result. This section analyzes the effect of the medical factors on the survival time of patients with the disease carried out at the Medical Centre of the University of West Virginia. Survival time of patients (ST), the level of blood urea nitrogen (BUN), serum calcium (CA), hemoglobin (HB), and the percentage of plasma cells (PC) in the bone marrow are considered. Data can be found in Krall et al. (1975). For simplicity, complete data points for males in the data set are used, where the sample size is 22.

Procedures are based on a copula chosen by the Kolmogorov-Smirnov type distance. We select a best copula within the class of the t copulas from the Kolmogorov-Smirnov type distance criterion and then check its adequacy using some goodness of fit tests. Finally, we explore the anticipated longest life span of a patient with the disease based on a large number of simulated data points generated by the copula. We start with finding the marginal distribution of the medical factors.

3.1 Marginal Distribution

Appropriate marginal distributions should be plugged in the t copula. Toward an optimal distribution for the data given, numerical model checking methods, such as the Kolmogorov-Smirnov and Anderson-Darling tests, were used to check if the distribution chosen is appropriate. The Kolmogorov-Smirnov test uses maximum difference between the empirical distribution and the theoretical distribution, defined as $sup_x \left\{ \left| \hat{F}(x) - F(x) \right| \right\}$ where $\hat{F}(x)$ is the empirical distribution. The Anderson-Darling test is the test that is more sensitive in the tails of the distribution than the Kolmogorov-Smirnov test. Best fitting marginal distributions for data are selected such that the distance function is minimized. Judging from those two criteria, we arrive at the distributions summarized in Table 2. Table 3 shows the correlations coefficients for the data.

Table 2. Distribution and parameter estimate for the multiple myeloma data

	Distribution	Parameter Estimate	Mean	Std. Dev.	Skewness	Kurtosis
ST	Lognormal	μ=2.47, σ=1.29	22.77	24.42	1.44	1.26
BUN	Lognormal	μ=3.28, σ=0.66	34.0	32.44	3.37	13.50
CA	Lognormal	μ=9.59, σ=1.26	10.32	1.62	1.37	2.14
HB	Lognormal	μ=14.79, σ=0.73	10.43	2.77	-0.14	-0.89
PC	Lognormal	μ=3.55, σ=0.65	42.59	27.45	0.94	0.01

Table 3. Correlation coefficient for the multiple myeloma data

	S	BUN	CA	HB	PC
S	1	-0.3454	-0.0196	0.3435	-0.2899
BUN	-0.3454	1	0.1219	-0.3624	-0.0599
CA	-0.0196	0.1219	1	0.2175	-0.0831
HB	0.3435	-0.3624	0.2175	1	-0.2985
PC	-0.2899	-0.0599	-0.0831	-0.2985	1

The distribution function gives the probability that a random variable is less than a given value, describing the parent distribution. The empirical distribution function is similar, the difference being that the empirical distribution is calculated by data, resembling the theoretical distribution that fits data. So the plot of $\hat{F}$ versus F should be close to each other if the distribution considered is legitimate for data. For example, Figure 3 shows that the empirical distribution for the survival time is fairly close to the specified distribution, the lognormal distribution. Hence the lognormal distribution appears to be an appropriate distributional model. This graphical method confirms the result from the numerical methods.

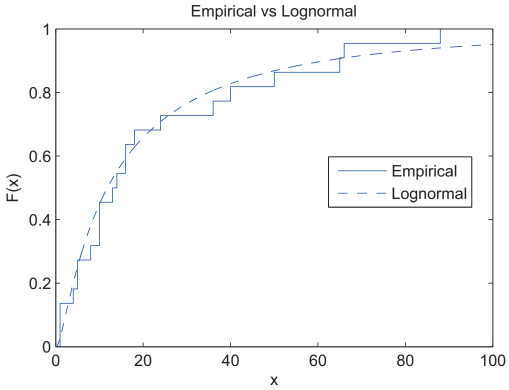


Figure 3. Plot of empirical vs lognormal for survival time

3.2 Copula Selection and Its Adequacy

Depending on the degrees of freedom, the t copulas specify different amounts of structural information, especially on tail behavior of distributions. Even with identical correlations, the dependence structure of distributions may be different depending on the choice of copulas. The t copula with low degrees of freedom captures the non-linear trends well, while the Gaussian copula or the t copula with high degrees of freedom fit well when the dependence is mostly linear. So a copula that well describes the structure of data should be used in practice. This section finds such a suitable copula to use and checks its statistical significance. The procedures are based on the empirical copula, $\hat{C}$, introduced in Deheuvels (1978), defined as

$$\hat{C}(u_1, \ldots, u_n) = \hat{H}(\hat{F}_1(x_1), \ldots, \hat{F}_n(x_n)), \tag{4}$$

where

$$\hat{H}(u_1, \ldots, u_n) = \frac{1}{n}\sum_{i=1}^{m} I\{U_{1,i} \le u_1, \ldots, U_{n,i} \le u_n\}$$

and $\hat{F}_i$'s are the usual empirical distributions that correspond to univariate marginal distributions. Similar to the empirical distribution that estimates an unknown distribution, the empirical copula obtained by data estimates a theoretical copula. The empirical copula in (4) actually provides approximate probabilities of the number of pairs $(u_1, \ldots, u_n)$ such that $u_1 \le u_{1(i)}, \ldots, u_n \le u_{n(i)}$, where $u_{(i)}$ denotes the order statistic. A best copula within a class of the t copulas is chosen such that the distance of the empirical copula $\hat{C}$ and the theoretical copula C is minimized (Durrleman et al., 2000). However, this does not mean that a copula chosen is a fitted model for data in an absolute sense. Statistical testing procedures should be conducted to decide whether the copula chosen is adequate for data. To this end, based on the Cramr-von Mises type statistic (Genest & Rmillard, 2008; Genest et al., 1993; Genest et al., 2009), define the process in x

$$D(x) = n\int_0^x \left\{\hat{C} - C\right\}^2 d\hat{C},$$

for $0 < x \le 1$. From this, define the simulated process,

$$D^*(x) = n\int_0^x \left\{\hat{C} - C^*\right\}^2 d\hat{C}, \tag{5}$$

where C^* is the simulated copula of C from which data are re-sampled. Under the null hypothesis that the copula C is valid for data, the empirical copula resembles the assumed copula. Therefore, comparison of D to a large number of simulated realizations from the process D^* in (5) will lead to goodness-of-fit tests for the model. Specifically, since the process $D(x)$ randomly fluctuates around zero under the null hypothesis, a distinguishably large value of D compared to the values of D^* would indicate model misspecification. Thus, $S = \sup_{0<x\leq 1} |D(x)|$ can be used as a numerical measure for the assessment of the model adequacy. A large value of S will lead to rejection of the null hypothesis. Let $S^* = \sup_{0<x\leq 1} |D^*(x)|$. Then, the $p-$value defined as $P(S \geq s)$, where s is the observed value of S, can be used as a measure of strength of the model adequacy, and this $p-$value is approximated by $P(S^* \geq s)$ which can be empirically estimated by the Monte Carlo simulation method, similar to Lee et al. (2008). We use this $p-$value as a numerical measure of how well a copula model chosen fits the data. Lower the $p-$value, the less likely the goodness of fit is.

4. Numerical Results

The goal of this study is to estimate the expected life span of a patient with the disease by finding a combination of the survival time and the medical factors that best summarizes it. We considered various cases, each of which contains the survival time denoted by ST. Specifically, we investigate the following four classes: 1. (ST, HB), (ST, CA), (ST, BUN), (ST, PC), 2. (ST, BUN, CA), (ST, BUN, HB), (ST, BUN, PC), (ST, CA, HB), (ST, CA, PC), (ST, HB, PC), 3. (ST, BUN, CA, HB), (ST, BUN, CA, PC), (ST, BUN, HB, PC), (ST, CA, HB, PC), and 4. (ST, BUN, CA, HB, PC). Classes 1, 2, 3, and 4 consider the association of two, three, four and all the factors, respectively. To find the best copula from the four classes, in the first step we performed the copula selection process mentioned in Section 3.2. Based on this, we arrived at the following cases as superior within their own class: (ST, HB), (ST, CA, HB), (ST, BUN, CA, HB), and (ST, BUN, CA, HB, PC). Results are summarized in Table 4. Note that to find the best t copula for each case, the degrees of freedom ranging from 1 to 30 was considered. The t copula with low degrees of freedom would be selected if the dependence of extreme values is detected.

Table 4. Copula selected and its $p-$value

Case	Copula	Distance/n	$p-$value
(ST, HB)	t copula, ν=6	0.0038	0.8580
(ST, CA, HB)	t copula, ν=16	0.0065	0.7250
(ST, BUN, CA, HB)	t copula, ν=6	0.0082	0.4590
(ST, BUN, CA, HB, PC)	t copula, ν=5	0.0192	0.3640

In the next step, we obtained the $p-$values using the goodness-of-fit test procedures described in Section 3.2 to check the adequacy of the copulas chosen. The estimated $p-$values obtained for (ST, HB), (ST, CA, HB), (ST, BUN, CA, HB), and (ST, BUN, CA, HB, PC) are presented in Table 4. Results are based on 2000 simulated realizations. Judging from the results, we conclude that the relationship between survival time and hemoglobin, (ST, HB), which is modeled by the t copula with the degrees of freedom 6 would most likely be distinguishable among others. Multiple myeloma has been staged by the method developed by Durie and Salmon (1975). From the staging method by Durie and Salmon (1975), it was found that the level of hemoglobin in the blood of a multiple myeloma patient is strongly associated with the tumor mass and thus is a strong indicator of the disease progress (Durie & Salmon, 1975; Kyle & Rajkumar, 2008). Our results are in agreement with this study.

Lastly, we examine the anticipated life expectancy of a patient with multiple myeloma from diagnosis until death by generating a large number of simulated survival times from the copula chosen. Specifically, based on the aforementioned results, in association with the marginal distributions described in Table 2, we re-sampled survival times of size 200,000 from the t copula (ν=6) with reference to hemoglobin only. Table 5 presents the estimated survival times at various percent confidence levels, from 5% (shortest) survival time to 95% (longest) survival time, in months, from diagnosis until death from multiple myeloma. Table 5 also shows the corresponding estimated hemoglobin levels. For example, the 95th percentile of the survival times indicates that survival times of a patient under treatment could extend to 99.13 months (approximately 8.3 years) and its corresponding hemoglobin level is 14.68. Figure 4 shows 2000 simulated realizations under the t copula with degrees of freedom 6. The positive effect of hemoglobin on survival time seems somewhat stronger than the negative effect in the scatter plot.

Table 5. Multiple myeloma life expectancy for (ST, HB), 200,000 samples

Percentile	5%	10%	30%	50%	70%	90%	95%
Estimated life expectancy	1.41	2.26	6.04	11.84	23.25	62.02	99.13
Estimated hemoglobin	7.52	7.53	6.67	7.89	9.51	10.65	14.68

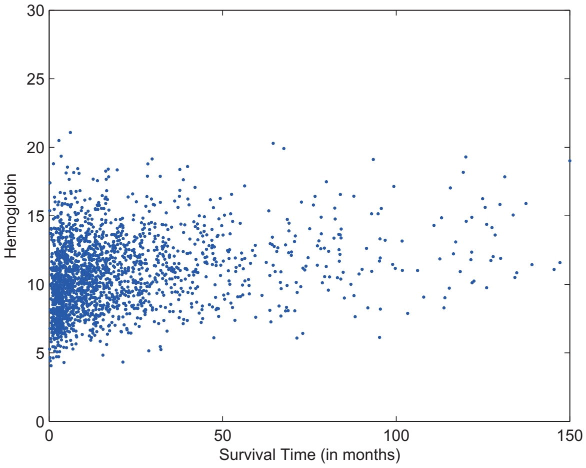


Figure 4. Survival time vs Hemoglobin, *t* copula with ν=6, 2000 simulated samples

5. Concluding Remarks

Multiple myeloma is a malignant disease that develops in the bone marrow. In most cases this disease develops in association with several medical factors that act together. A copula provides a convenient tool in the analysis of multivariate data that represent such medical factors. It was found that the relationship between survival time and hemoglobin which is modeled by the *t* copula with the degrees of freedom 6 would be most distinguishable. Some goodness-of-fit tests based on the simulated process were then performed to check the adequacy of the copula model. Based on the copula, we simulated a large number of simulated realizations of the survival time from which the anticipated life span of a patient with the disease was calculated.

References

Breymann, W., Dias, A., & Embrechts, P. (2003). Dependence Structures for Multivariate High- Frequency Data in Finance. *Quantitative Finance, 3*(1), 1-16. http://dx.doi.org/10.1080/713666155

Collect, D. (1999). *Modeling Survival Data in Medical Research*. Chapman.

Deheuvels, P. (1978). Caracterisation complete des lois extremes multivariees et de la convergence des types extremes. *Publications de l'Institut de Statistique de l'Universite de Paris, 23*, 1-36.

Durie, B. G., & Salmon, S. E. (1975). A clinical staging system for multiple myeloma. Correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival. *Cancer, 36*(3), 842-854. http://dx.doi.org/10.1002/1097-0142(197509)36:<842::AID-CNCR2820360303>3.0.CO;2-U

Durrleman, V., Nikeghbail, A., & Roncalli, T. (2000). Which copula is the right one? Credit Lyonnais. Retrieved from http://ssm.com/abstract=1032545

Embrechts, P., Lindskog, F., & McNeil, A. (2003). Modelling Dependence with Copulas and Applications to Risk Management. In S. Rachev (Ed.), *Handbook of Heavy Tailed Distributions in Finance* (Chapter 8, 329-384), Elsevier. http://dx.doi.org/10.1016/B978-044450896-6.50010-8

Embrechts, P., McNeil, A., & Straumann, D. (2002). Correlation and Dependency in Risk Management: Properties and Pitfall. In M. Dempster (Ed.), *In Risk Management: Value at Risk and Beyond* (pp. 176-223). Cambridge: Cambridge University Press.

Frees, M. J., & Valdez, E. (1998). Understanding relationships using copulas. *North American Actuarial Journal, 2*, 1-25.

Genest, C., & Rivest, L. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of American Statistical Association, 88*, 1034-1043. http://dx.doi.org/10.1080/01621459.1993.10476372

Genest, C., & Rmillard, B. (2008). Validity of the Parametric Bootstrap for Goodness-of-fit Testing in Semiparametric Models. *Annales l'Institut Henri-Poincare, 44*, 1096-1127.

Genest, C., Rmillard, B., & Beaudoin, D. (2009). Goodness-of-fit Tests for Copulas: A Review and a Power Study.

Insurance: Mathematics and Economics, 44, 199-213. http://dx.doi.org/10.1016/j.insmatheco.2007.10.005

Krall, J. M, Uthoff, V. A., & Harley, J. B. (1975). A step-up procedure for selecting variables associated with survival. *Biometrics, 31*, 49-51. http://dx.doi.org/10.2307/2529709

Kumar, P., & Shoukri, M. M. (2008). Evaluating Aortic Stenosis using the Archimedean copula methodology. *Journal of Data Science, 6*, 173-187.

Kyle, R. A., & Rajkumar, S. V. (2008). Multiple myeloma. *Blood, 111*(6), 2962-2972. http://dx.doi.org/10.1182/blood-2007-10-078022

Lee, S., Lee, E.-J., & Omolo, B. O. (2008). Using integrated weighted survival difference for the two-sample censored data problem. *Computational Statistics and Data Analysis, 52*, 4410-4416. http://dx.doi.org/10.1016/j.csda.2008.02.022

Melchiori, M. R. (2003). Which Archimedean copula is the right one? *Yield Curve, 37*, 1-20.

Multiple Myeloma: Disease Overview. (2010). Multiple Myeloma Research Foundation. Retrieved from www.themmrf.org

Nelsen, R. B. (1999). *An introduction to copulas*. Springer. http://dx.doi.org/10.1007/978-1-4757-3076-0

Schweizer, B., & Wolff, E. F. (1981). On the Nonparametric Measures of Depdendence for Random Variables. *Annals of Statistics, 9*, 879-885. http://dx.doi.org/10.1214/aos/1176345528

Sklar, A. (1959). *Functions de Repartition a n Dimensions et leurs Merges*. Publication of the Institute of Statistics, University of Paris 8, 229-231.

Zheng, M., & Klein, J. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika, 82*, 127-138. http://dx.doi.org/10.1093/biomet/82.1.127

International Journal of Statistics and Probability; Vol. 2, No. 1; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Quasi Maximum Likelihood Approach to Statistical Inference on a Nonstationary Multivariate ARFIMA Process

Amadou Kamagaté[1,2] & Ouagnina Hili[2]

[1] Nangui Abrogoua University, Abidjan, Côte d'Ivoire

[2] Laboratory of Mathematics and New Technologies of Information, National Polytechnic Institute Félix Houphouët-Boigny, Yamoussoukro, Côte d'Ivoire

Correspondence: Ouagnina Hili, Laboratory of Mathematics and New Technologies of Information, National Polytechnic Institute Félix Houphouët-Boigny, Yamoussoukro, Yamoussoukro BP 1093, Côte d'Ivoire. E-mail: o_hili@yahoo.fr

Received: September 3, 2012 Accepted: December 20, 2012 Online Published: January 16, 2013

doi:10.5539/ijsp.v2n1p53 URL: http://dx.doi.org/10.5539/ijsp.v2n1p53

Abstract

In this Note, we estimate the parameters of a nonstationary multivariate ARFIMA (AutoRegressive Fractionally Integrated Moving Average) process by the quasi likelihood approach. Then, we define the pseudo spectral density of the process. Under some assumptions, we establish Consistency and Asymptotic normality.

Keywords: quasi maximum likelihood, long memory, multivariate time series, nonstationary ARFIMA process

1. Introduction

We have proposed in this note an another estimation method for an Auto Regressive Fractionally Integrated Moving Average (ARFIMA) process. Granger and Joyeux (1980) have proposed a class of process whose d is a real value in ARIMA(p, d, q) process of (Box & Jenkins, 1976). This model is well known to have many applications in financial statistics. Granger (1980) introduced the ARFIMA model where $d \in (0, 1/2)$. For $d \in (-1/2, 1/2)$, the multivariate ARFIMA process is stationary and has a spectral density as in (Hosoya, 1996).

For $d > 1/2$, we define a pseudo spectral density in (5) as in (Nielsen, 2009). A number of authors estimated the parameters in the case of nonstationary processes with respect to his method. Also, for a univariate process, Phillips and Shimotsu (2004) showed that the Whittle estimation of a nonstationary process is inconsistent as $d > 1$ but Shao (2009) showed that extended Whittle estimation is consistent for $d > 1$. The resulting estimate as in (Shao, 2009) is asymptotically normal and is more efficient than the tapered Whittle estimate as in (Velasco, 1999a). For the multivariate process, Nielsen (2009) generalized the works of (Shao, 2009). He showed that the Extended Multivariate Local Whittle Estimation (ExtMLWE) improves the Multivariate Local Whittle Estimation (MLWE) used by Shimotsu (2007).

The present note estimates the parameters by the quasi likelihood method for estimation of a nonstationary multivariate process. We use the general form of the spectral density as in (Hosoya, 1996) and then we extend his results to a nonstationary case of the process by using the extended discrete Fourier transform and the periodogram as in (Nielsen, 2009). The works of this paper have permitted to establish consistency and asymptotic normality for a nonstationary multivariate fractionally integrated process.

The L^p-norm of a complex-valued function g on $(-\pi; \pi]$ is denoted by $\|g\|_p$ and it is defined by $[\int_{-\pi}^{\pi} |g(\lambda)|^p d\lambda]^{1/p}$. Moreover, we use positive constants c_j where $j = 1, 2, \ldots$ in this note. The notations A_{ij} and B_i or A and B denote in order a matrix or a column vector.

We organize this note as follows. The ARFIMA model and the quasi likelihood function are introduced in section 2. Section 3 constitutes the main result of this note. The quasi maximum likelihood estimation method and the asymptotic properties of the estimate are discussed in this section. All proofs are gathered in section 4. We conclude in section 5 by some simulations.

2. The Model and the Quasi Maximum Likelihood Function

Granger and Joyeux (1980) and Hosking (1981) have proposed the ARFIMA (p, d, q) models to define a time series which presents a character of short or long memory following d. The model of multiple ARFIMA processes was introduced by Sowell (1987; 1989). We consider a m-dimensional ARFIMA nonstationary process $\{y_1(t), \dots, y_m(t)\}$ following $d_0 > 1/2$ which is generated by

$$\begin{pmatrix} A_{1,1}(L) & \dots & A_{1,m}(L) \\ \vdots & & \vdots \\ A_{m,1}(L) & \dots & A_{m,m}(L) \end{pmatrix} \begin{pmatrix} (1-L)^{d_{1,0}} \\ \vdots \\ (1-L)^{d_{m,0}} \end{pmatrix} \begin{pmatrix} y_1(t) \\ \vdots \\ y_m(t) \end{pmatrix} = \begin{pmatrix} B_{1,1}(L) & \dots & B_{1,m}(L) \\ \vdots & & \vdots \\ B_{m,1}(L) & \dots & B_{m,m}(L) \end{pmatrix} \begin{pmatrix} \varepsilon_1(t) \\ \vdots \\ \varepsilon_m(t) \end{pmatrix}, \tag{1}$$

$d_{i,0} = d^*_{i,0} + r$ with $d^*_{i,0} \in (-1/2, 1/2)$ and $r \geq 1$ is a positive integer.

For all $r \geq 1$, (1) becomes

$$\begin{pmatrix} A_{1,1}(L) & \dots & A_{1,m}(L) \\ \vdots & & \vdots \\ A_{m,1}(L) & \dots & A_{m,m}(L) \end{pmatrix} \begin{pmatrix} (1-L)^{d^*_{1,0}} \\ \vdots \\ (1-L)^{d^*_{m,0}} \end{pmatrix} \begin{pmatrix} (1-L)^r \\ \vdots \\ (1-L)^r \end{pmatrix} \begin{pmatrix} y_1(t) \\ \vdots \\ y_m(t) \end{pmatrix} = \begin{pmatrix} B_{1,1}(L) & \dots & B_{1,m}(L) \\ \vdots & & \vdots \\ B_{m,1}(L) & \dots & B_{m,m}(L) \end{pmatrix} \begin{pmatrix} \varepsilon_1(t) \\ \vdots \\ \varepsilon_m(t) \end{pmatrix}, \tag{2}$$

where $\{\varepsilon_1(t), \dots, \varepsilon_m(t)\}$ have a distribution law with mean zero and $\text{cov}\{\varepsilon_i(t), \varepsilon_j(s)\} = \delta(t,s)K_{ij},\ i,j = 1, \dots, m$, $K = \{K_{ij}\}$ supposed positive definite, $r - 1/2 < d_{i,0} < r + 1/2$ $(i = 1, \dots, m)$ with $d_{i,0} = d^*_{i,0} + r$ and L is the backshift operator defined by $Ly_t = y_{t-1}$.

Letting $(1-L)^r y(t) = X(t)$, (2) becomes

$$\begin{pmatrix} A_{1,1}(L) & \dots & A_{1,m}(L) \\ \vdots & & \vdots \\ A_{m,1}(L) & \dots & A_{m,m}(L) \end{pmatrix} \begin{pmatrix} (1-L)^{d^*_{1,0}} \\ \vdots \\ (1-L)^{d^*_{m,0}} \end{pmatrix} \begin{pmatrix} X_1(t) \\ \vdots \\ X_m(t) \end{pmatrix} = \begin{pmatrix} B_{1,1}(L) & \dots & B_{1,m}(L) \\ \vdots & & \vdots \\ B_{m,1}(L) & \dots & B_{m,m}(L) \end{pmatrix} \begin{pmatrix} \varepsilon_1(t) \\ \vdots \\ \varepsilon_m(t) \end{pmatrix}, \tag{3}$$

where $A_{i,j}(L) = \sum_{k=0}^{p} A_{i,j}(k)L^k$ and $B_{i,j}(L) = \sum_{k=0}^{q} B_{i,j}(k)L^k$ such as $A(0) = I_m$ and the zeros of $\det A(z)$ and $\det B(z)$ are outside the unit circle.

Also, we assume here-after that $\{X_t\}$ is a second-order stationary ARFIMA $(p, d^*_{i,0}, q)$ process. Odaki (1993) and Hosking (1981) have showed that $\{X_t\}$ is invertible and stationary respectively as long as $d^*_{i,0} > -1$ and $d^*_{i,0} < 1/2$.

So, the process is short memory if $-1/2 < d^*_{i,0} < 0$ and long memory if $0 < d^*_{i,0} < 1/2$. Let θ be the vector whose components consist of $d^*_1, \dots, d^*_m$, $A_{1,1}(k), A_{1,2}(k), \dots, A_{m,m}(k)$, $B_{1,1}(k), B_{1,2}(k), \dots, B_{m,m}(k)$. Suppose that the parameter space Θ is a compact subset of $\mathbb{R}^{m(p+q+1)}$.

Considering the conditions on the polynomials, the process is invertible and causal. (3) can be rewritten as an infinite moving average (MA(∞)) representation:

$$\begin{pmatrix} X_1(t) \\ \vdots \\ X_m(t) \end{pmatrix} = \begin{pmatrix} G_{1,1}(L,\theta) & \dots & G_{1,m}(L,\theta) \\ \vdots & & \vdots \\ G_{m,1}(L,\theta) & \dots & G_{m,m}(L,\theta) \end{pmatrix} \begin{pmatrix} \varepsilon_1(t) \\ \vdots \\ \varepsilon_m(t) \end{pmatrix}, \tag{4}$$

where $\varepsilon(t)$ is a white noise and the infinite polynomials $G_{i,j} = \sum_{k=0}^{\infty} G_{i,j}(k,\theta)L^k$ are determined in (4) in view of the following relationship

$$(1-L)^d \quad = \quad 1 + \sum_{k=0}^{\infty} \frac{\Gamma(k-d)}{k!\Gamma(-d)} L^k.$$

We assume that the coefficients matrices $G(j,\theta)$ satisfy $\sum_{j=0}^{\infty} \text{tr}\, \text{G}(\text{j},\theta)\text{K}\text{G}(\text{j},\theta)^* < \infty$ and the process has a $m \times m$ pseudo spectral density matrix determined by

$$f_\theta(\lambda) \quad := |1 - \exp(i\lambda)|^{-2r} \tfrac{1}{2\pi}\Big[\Big(\textstyle\sum_{j=0}^{\infty} G(j,\theta)\exp(i\lambda j)\Big)K\Big(\sum_{j=0}^{\infty} G(j,\theta)\exp(i\lambda j)\Big)^*\Big]. \tag{5}$$

Now, we define the extended discrete Fourier transform (EDFT) and the extended periodogram matrix of $y(t)$ evaluated at the Fourier frequencies $\lambda_j = \frac{2\pi j}{n}$, where $j = 1, \dots, n$ by

$$w_j(d_0) = w(\lambda_j, d_0) = w^y(\lambda_j) + c(\lambda_j, d_0) \tag{6}$$

$$I_j(d_0) = I(\lambda_j, d_0) = w(\lambda_j, d_0)w^*(\lambda_j, d_0) \tag{7}$$

where $w^y(\lambda_j)$ is the DFT defined as

$$w^y(\lambda_j) = \frac{1}{\sqrt{(2\pi n)}}\sum_{t=1}^{n} y(t)e^{it\lambda_j}, \tag{8}$$

and the correction term for the ith element $c_i(\lambda_j, d_0)$ takes on constant values on the intervals $d_{i,0} \in D_r := [r - 1/2, r + 1/2)$, $r \in \mathbb{N}^*$, $i = 1, \ldots, m$ and is defined by

$$c_i(\lambda_j, d_0) = \begin{cases} 0, & \text{if } d_{i,0} \in D^0 = [-1/2, 1/2), \\ e^{i\lambda_j}\sum_{l=1}^{r}(1 - e^{i\lambda_j})^{-l}Z_{i,l}, & \text{if } d_{i,0} \in D_r \text{ for } r \geq 1. \end{cases} \tag{9}$$

where

$$Z_{i0} = w_{iy}(0) = \frac{1}{\sqrt{(2\pi n)}}\sum_{t=1}^{n} y_i(t) \tag{10}$$

$$Z_{il} = \frac{1}{\sqrt{(2\pi n)}}\{(1 - L)^{l-1}y_i(n) - (1 - L)^{l-1}y_i(0)\}, l = 1, 2, \ldots, r. \tag{11}$$

The $w(\lambda_j, d_0)$ for $j = 1, \ldots, n$ is a multivariate normal distribution and they are approximately independent when n is large, the probability density function

$$\frac{1}{\pi^2\sqrt{\det f_\theta(\lambda_j)}}\exp\Big(\frac{-1}{2}\mathrm{tr}\Big(f_\theta^{-1}(\lambda_j)w(\lambda_j, d_0)w(\lambda_j, d_0)^*\Big)\Big),\ j = 1, \ldots, n, \tag{12}$$

(see Hannan & Edward, 1970, pp. 224-225). An approximate log-likelihood $L_n(\theta)$ following these observations $y(1), \ldots, y(n)$ is defined by the following expression

$$L_n(\theta) = -\sum_{j=1}^{n}\Big(\log\det f_\theta(\lambda_j) + \mathrm{tr}(f_\theta^{-1}(\lambda_j)I_n(\lambda_j, d_0)\Big). \tag{13}$$

The integral form of (13) is

$$\overline{L}_n(\theta) = -n\Big[\int_{-\pi}^{+\pi}\log\det f_\theta(\lambda)d\lambda + \int_{-\pi}^{+\pi}\mathrm{tr}\Big(f_\theta^{-1}(\lambda)I_n(\lambda, d_0)\Big)d\lambda\Big]. \tag{14}$$

The function $\overline{L}_n(\theta)$ is termed the quasi-log-likelihood function.

3. Quasi-maximum-likelihood Estimation of the Parameters

Assume throughout that $\partial\int_{-\pi}^{+\pi}\log\det f_\theta(\lambda)d\lambda/\partial\theta_j$ exists and at almost all points of θ, $\partial f_\theta^{-1}(\lambda)/\partial\theta_j$ exists following θ. We denote $H_j(\theta) = \partial\int_{-\pi}^{+\pi}\log\det f_\theta(\lambda)d\lambda/\partial\theta_j$ and $h_j(\lambda, \theta) = \partial f_\theta^{-1}(\lambda)/\partial\theta_j$. For $j = 1, \ldots, p + q + 1$, the notations $H_j(\theta)$ and $\mathrm{tr}\{h_j(\lambda, \theta)f(\lambda)\}$ represent in order the jth elements of $H(\theta)$ and $\mathrm{tr}\{h_\theta(\lambda)f(\lambda)\}$. Suppose that

$$S_{nj}(\theta) = H_j(\theta) + \int_{-\pi}^{\pi}\mathrm{tr}\{h_j(\lambda, \theta)I_n(\lambda, d_0)\}d\lambda, \quad j = 1, \ldots, p + q + 1 \tag{15}$$

and denote by $S_n(\theta)$ the vector $\{S_{nj}(\theta)\}$, where $I_n(\lambda, d_0)$ is the extended periodogram matrix defined similarly to (7). The quasi maximum likelihood estimate (QMLE) of θ is defined by the value $\widetilde{\theta}_n$ such that $S_n(\widetilde{\theta}_n) = 0$.

Suppose that $\{\varepsilon(t)\}$ is fourth-order stationary and that

$$\sum_{t_1, t_2, t_3=-\infty}^{\infty}|\widetilde{Q}^{\varepsilon}_{\beta_1,\ldots,\beta_4}(t_1, t_2, t_3)| < \infty,$$

where $\widetilde{Q}^{\varepsilon}_{\beta_1,\ldots,\beta_4}(\lambda_1, \lambda_2, \lambda_3)$ is the joint fourth cumulant of $\varepsilon_{\beta_1}(t), \varepsilon_{\beta_2}(t+t_1), \varepsilon_{\beta_2}(t+t_2), \varepsilon_{\beta_3}(t+t_3)$ moreover $Q^{\varepsilon}_{\beta_1,\ldots,\beta_4}(\lambda_1, \lambda_2, \lambda_3)$ is a fourth-order spectral density of $\{\varepsilon(t)\}$ defined by

$$Q^{\varepsilon}_{\beta_1,\ldots,\beta_4}(\lambda_1, \lambda_2, \lambda_3) = \frac{1}{(2\pi)^3}\sum_{t_1, t_2, t_3=-\infty}^{\infty}\exp\{-i(\lambda_1 t_1 + \lambda_2 t_2 + \lambda_3 t_3)\}\widetilde{Q}^{\varepsilon}_{\beta_1,\ldots,\beta_4}(t_1, t_2, t_3).$$

We assumed Assumption A throughout the note.

Assumption A

(A1) There exists $\epsilon > 0$ such that for any $t < t_1 \leq t_2 \leq t_3 \leq t_4$ and for each β_1,β_2, $Var\Big[E\{\varepsilon_{\beta_1}(t_1)\varepsilon_{\beta_2}(t_2)|\mathcal{B}(t)\} - \delta(t_1-t_2,0)K_{\beta_1\beta_2}\Big] = O\{(t_1-t)^{-2-\epsilon}\}$, $\epsilon > 0$ and $E\Big|E\{\varepsilon_{\beta_1}(t_1)\varepsilon_{\beta_2}(t_2)\varepsilon_{\beta_3}(t_3)\varepsilon_{\beta_4}(t_4)|\mathcal{B}(t)\} - E\{\varepsilon_{\beta_1}(t_1)\varepsilon_{\beta_2}(t_2)\varepsilon_{\beta_3}(t_3)\varepsilon_{\beta_4}(t_4)\}\Big| = O\{(t_1-t)^{-1-\epsilon}\}$, uniformly in t, where $\mathcal{B}(t)$ is the σ-field generated by $\{\varepsilon(s), s \leq t\}$.

(A2) For any $\epsilon > 0$ and for any integer $M \geq 0$, there exists $B_\epsilon > 0$ such that $E[T(n,s)^2 I\{T(n,s) > B_\epsilon\}] < \epsilon$, uniformly in $n,\ s$, where I implies the indicator function and

$$T(n,s) \quad = \quad \Big[\sum_{\alpha,\beta=1}^{m}\sum_{r=0}^{M}\Big\{\sum_{t=1}^{n}(\varepsilon_\alpha(t+s)\varepsilon_\beta(t+s+r) - K_{\alpha\beta}\delta(0,r))/n^{\frac{1}{2}}\Big\}^2\Big]^{\frac{1}{2}}.$$

(A3) Each component $Q^{\varepsilon}_{\beta_1,\ldots,\beta_4}(\lambda_1,\lambda_2,\lambda_3)$ is uniformly γ-lipschitz for some $\gamma > 0$, namely

$$\Big|Q^{\varepsilon}_{\beta_1..\beta_4}(\lambda_1+\epsilon_1,\lambda_2+\epsilon_2,\lambda_3+\epsilon_3) - Q^{\varepsilon}_{\beta_1,\ldots,\beta_4}(\lambda_1,\lambda_2,\lambda_3)\Big| \quad < \quad \{\max_i |\epsilon_i|\}^\gamma,$$

uniformly in $\lambda_1,\lambda_2,\lambda_3$.

To establish the consistence and the asymptotic normality, we assume the following conditions.

Assumption B

(B1) The process observed $y(t)$ has a pseudo spectral density

$$f_\theta(\lambda) \quad := |1-e^{i\lambda}|^{-2r}\tfrac{1}{2\pi}k(\lambda)Kk(\lambda)^*$$

that satisfies

1) $\int_{-\pi}^{\pi}|k_{\alpha\beta}(\lambda)|^{2u}d\lambda < \infty$, for u such that $1 < u \leq 2$, $\alpha,\beta = 1,\ldots,m$, where $k(\lambda) = \sum_{j=0}^{\infty}G(j,\theta)\exp(i\lambda j)$, $k_{\alpha\beta}$ the (α,β)th element of the matrix $k(\lambda)$ and K the covariance matrix of $\{\varepsilon(t)\}$.

2) Let $c > 0$, following

$$\sup_{|\nu|<\epsilon}\max_{\alpha,\beta}\Big\|\Big[f_\theta^{-1}(.)\{f_\theta(.) - f_\theta(.-\nu)\}\Big]_{\alpha\beta}\Big\|_u \quad = \quad O(\epsilon^c), \tag{16}$$

(B2) For any $\epsilon > 0$, there exists $a > 0$ and Hermitian-valued bounded functions $\widetilde{h}_j$ and $\overline{h}_j$ such that if $|\theta_1 - \theta| < a$,

$$\widetilde{h}_j \quad \leq \quad h_j(\lambda,\theta_1) \leq \overline{h}_j$$

and

$$\max_{\alpha,\beta}\|[\{\overline{h}_j - \widetilde{h}_j\}f_\theta(.)]_{\alpha,\beta}\|_\upsilon \quad < \quad \epsilon, \tag{17}$$

where A $\leq$ B denotes that B $-$ A is positive definite and $\upsilon = (u-1)/u$ for u defined in (16).

(B3) $V_j(\theta) = 0$ for all j at $\theta = \theta_0$ where $\theta_0 \in \Theta$ and $V_j(\theta) \equiv H_j(\theta) + \int_{-\pi}^{\pi}\mathrm{tr}\{h_j(\lambda,\theta)f_\theta(\lambda)\}d\lambda$.

(B4) $H_j(\theta)$ is continuous on Θ.

Theorem 1 *Under Assumption B, the QML estimate $\widetilde{\theta}_n$ tends to θ_0 in probability.*

Assumption C The pair $\{g^{(1)}(\lambda), k^{(1)}(\lambda)\}$ of complex functions holds:

(C1) For some $\gamma,\epsilon > 0$, $|g^{(1)}(\lambda) - g^{(1)}(\lambda+\epsilon)| < |\epsilon|^\gamma$ uniformly in λ.

(C2) There exists $u > 1$ such that $\int_{-\pi}^{\pi}|k^{(1)}(\lambda)|^{2u}d\lambda < \infty$.

(C3) There exists $\gamma_1 > 0$ such that

$$\sup_{|\epsilon|\leq\epsilon_1}\|g^{(1)}(.)|k^{(1)}(.+\epsilon) - k^{(1)}(.)|^2\|_2 = O(|\epsilon_1|^{\gamma_1}).$$

(C4) $\|g^{(1)}|k^{(1)}|^2\|_2 < \infty$.

Assumption D

(D1) For some $c > 1/2$, the relation (16) satisfies.

(D2) $\lim_{a\to 0}\sup_{|\theta-\theta_0|\le a}\|[\{h_j(.,\theta)-h_j(.,\theta_0)\}f_\theta(.)]_{\alpha\beta}\| < \infty$, for some $a > 0$, $j = 1,\dots,p{+}q{+}1$ and $\alpha,\beta = 1,\dots,m$.

(D3) Given $\epsilon > 0$, there exists an integer $m(\epsilon)$, a partition $U^1(a),\dots,U^{m(\epsilon)}(a)$ of the ball in Θ, with center θ_0 and radius a, and bounded Hermitian-matrix-valued $\widetilde{h}^i_j(\lambda)$; $\overline{h}^i_j(\lambda)$ such that, for all considerably small a and for all j; $\widetilde{h}^i_j(\lambda) \le h(\lambda,\theta) \le \overline{h}^i_j(\lambda)$ if $\theta \subset U^i(a)$ and also

$$\|[k^*(\widetilde{h}^i_j - h^0_j)k]_{\beta_1\beta_2}\|_v \le \epsilon a, \qquad \|[k^*(\overline{h}^i_j - h^0_j)k]_{\beta_1\beta_2}\|_v \le \epsilon a, \tag{18}$$

where $h^0_j = h_j(.,\theta_0)$ and * denotes the transpose.

(D4) $|V(\theta)| \ge \alpha_1|\theta-\theta_0|$ for some $\alpha_1 > 0$ in neighborhood of θ_0.

(D5) Assumption C holds for all pairs $\{\overline{h}_{j\alpha_2\alpha_1}; k_{\alpha\beta}\}$, $\{\widetilde{h}_{j\alpha_2\alpha_1}; k_{\alpha\beta}\}$ and $\{h_{j\alpha_2\alpha_1}; k_{\alpha\beta}\}$ where $\alpha = \alpha_1$ or α_2 and $1 \le \beta \le m$.

Theorem 2 *Under Assumptions B and D,* $\sqrt{n}(\widetilde{\theta}_n - \theta_0) \longrightarrow^{\mathcal{D}} N(0, W^{-1}U(W^{-1})^*)$ *if* V *is differentiable at* $\theta = \theta_0$ *and* $W_{ij} = \partial V_i/\partial\theta_j$ *evaluated at* $\theta = \theta_0$ *and* $U_{j,l}$ *is defined as*

$$\begin{aligned} U_{jl} &= 4\pi\int_{-\pi}^{\pi}\mathrm{tr}\{h_j(\lambda,\theta)f_j(\lambda)h_l(\lambda,\theta)f_l(\lambda)\}d\lambda 2\pi\sum_{\beta_1,\dots,\beta_4=1}^{m}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\Big[k^*(\lambda_1)h_j(\lambda_1,\theta_0)k(\lambda_1)\Big]_{\beta_1\beta_2} \\ &\quad\times\Big[k^*(\lambda_2)h_l(\lambda_2,\theta_0)k(\lambda_2)\Big]_{\beta_3\beta_4}Q^{\varepsilon}_{\beta_1,\dots,\beta_4}(\lambda_1,\lambda_2,-\lambda_2)d\lambda_1 d\lambda_2. \end{aligned} \tag{19}$$

4. Proofs

Proof of Theorem 1. Suppose that $\epsilon > 0$ and $B(a(\theta)) = \{\theta : |\theta_1 - \theta| < a(\theta)\}$. Moreover $\overline{h}_j(\lambda)$ and $\widetilde{h}_j(\lambda)$ are functions which verify (17) for ϵ. $a = a(\theta)$. We have

$$\begin{aligned} \sup_{B(a(\theta))}|S_{nj}(\theta_1) - S_{nj}(\theta)| &\le |H_j(\theta_1) - H_j(\theta)|\int_{-\pi}^{\pi}\mathrm{tr}\Big[(\overline{h}_j(\lambda) - \widetilde{h}_j(\lambda))f_\theta(\lambda)\Big]d\lambda \\ &\quad+\int_{-\pi}^{\pi}\mathrm{tr}\Big\{\overline{h}_j(\lambda) - \widetilde{h}_j(\lambda)\Big\}\Big\{I_n(\lambda,d_0) - E(I_n(\lambda,d_0))\Big\}d\lambda \\ &\le c_1\epsilon_1 + o_p(1), \end{aligned}$$

for $j = 1,\dots,p+q+1$. Then

$$\sup_{B(a(\theta))}|S_n(\theta_1) - S_n(\theta)| \le c_1(p+q+1)\epsilon_1 + o_p(1).$$

Since $V(\theta)$ is continuous due to Assumption B, then for an open neighborhood N of θ_0 there exists $\epsilon_2 > 0$ such that $\inf_{\Theta-N}|V(\theta)| > \epsilon_2$. We assume that $B_j = B(a(\theta_j))$, $j = 1,\dots,k$ is an open finite of $\Theta - N$. Then,

$$\begin{aligned} \inf_{\Theta-N}|S_n(\theta)| &\ge \inf_j|V(\theta_j)| - \sup_{B_j}|S_n(\theta)| + \sup_j|S_n(\theta_j) - V(\theta_j)| \\ &\ge \epsilon_2 - c_1(p+q+1)\epsilon_1 + o_p(1), \end{aligned}$$

since $\sup_j|S_n(\theta_j) - V(\theta_j)| \longrightarrow 0$ in probability. We choose ϵ_1 to have $\epsilon_2 - c_1(p+q+1)\epsilon_1 > 0$ and $\epsilon = \epsilon_2 - c_1(p+q+1)\epsilon_1 > 0$.

The next lemma constitutes the main part of Theorem 2.

Lemma 1 $\sqrt{n}\{S_n(\theta_0) + V(\widetilde{\theta}_n)\} \longrightarrow^{p}_{n\longrightarrow\infty} 0$ *if* $\sqrt{n}\{S_n(\widetilde{\theta}_n)\} \longrightarrow^{p} 0$ *and* $Pr\{|\widetilde{\theta}_n - \theta_0| \le b_0\} \longrightarrow 0$ *as* $n \longrightarrow \infty$, *for* $b_0 > 0$ *and considerably small.*

Proof of Lemma 1. We shall show that in probability

$$\sup_{|\theta-\theta_0|\le b_0}|S_n(\theta) - S_n(\theta_0) - V(\theta)|/\{n^{-1/2} + |V(\theta)|\} \longrightarrow^{p}_{n\longrightarrow\infty} 0.$$

This lemma is as in (Huber, 1967). For simplify the expressions, we define $h_j(\lambda;\theta) - h_j(\lambda;\theta_0)$ by $h - h^0$ and set $T_n(g)$ defined as

$$T_n(g) \quad = \quad \int_{-\pi}^{\pi} \mathrm{tr}\Big\{(g - h^0)(I_n(\lambda, d_0) - E(I_n(\lambda, d_0)))\Big\}d\lambda.$$

Also set $b_0 = 1$. Notice that in the following inequality

$$\sup_{|\theta|\le 1}\left|\int_{-\pi}^{\pi}\mathrm{tr}\Big\{(h-h^0)(I_n(\lambda,d_0)-f_\theta)\Big\}d\lambda\right| \quad\le\quad \sup_{|\theta|\le 1}|T_n(h)| + \sup_{|\theta|\le 1}\left|\int_{-\pi}^{\pi}\mathrm{tr}\Big\{(h-h^0)(E(I_n(\lambda,d_0))-f_\theta)\Big\}d\lambda\right|.$$

As in (Hosoya, 1997), the second member on the right-hand side of the preceding inequality is bounded by

$$\sup_{|\theta|\le 1}\left|\int_{-\pi}^{\pi}\mathrm{tr}\{(h-h^0)(E(I_n(\lambda,d_0))-f_\theta)\}d\lambda\right| \le C_1\sum_{\alpha,\beta=1}^{m}\|\{(h-h^0)f_\theta\}_{\alpha,\beta}\|_v \times \|\{f_\theta^{-1}(E(I_n(\lambda,d_0))-f_\theta)\}_{\beta,\alpha}\|_u.$$

Under (B1): $\|\{f_\theta^{-1}(E(I_n(\lambda,d_0))-f_\theta)\}_{\beta,\alpha}\|_u = O(n^{-c})$ and under (D2): $\sup_{|\theta-\theta_0|\le r}\|\{(h-h^0)f_\theta\}_{\alpha,\beta}\|_v < C_2$ where $C_2 > 0$. Now, we shall show that, as $n \longrightarrow \infty$

$$\sup_{|\theta|\le 1}|T_n(h)|/\{n^{-1/2} + |V(\theta)|\} \longrightarrow 0.$$

Choose l_0 such that $\frac{n}{2} < 4^{l_0+1} < n$ and $B(l)$ is the ball with center θ_0 and radius 2^{-l}, $l = 0, 1, \ldots, l_0$ and $A(l)$ denotes $B(l-1) - B(l)$. Suppose that $\epsilon > 0$, let $U^1, \ldots, U^m$ be a partition of $A(l)$ which satisfies (18) of (D3) for $\epsilon' > 0$. Suppose that

$$Q(2^{-l}) \quad = \quad \max_{\beta_1\beta_2}\{|\|[k^*(\overline{h}^i - h^0)k]_{\beta_1\beta_2}\|_v + |\|[k^*(\widetilde{h}^i - h^0)k]_{\beta_1\beta_2}\|_v\}.$$

As in (Hosoya, 1997),

$$T_n(h) \quad\le\quad T_n(\overline{h}^i) + \int_{-\pi}^{\pi}\mathrm{tr}\{(\overline{h}^i - \widetilde{h}^i)E(I_n(\lambda,d_0))\}d\lambda,$$

whereas, for $\theta \in U^i$, it follows from Assumption (D3) that

$$\begin{aligned}\left|\int_{-\pi}^{\pi}\mathrm{tr}\{(\overline{h}^i - \widetilde{h}^i)E(I_n(\lambda,d_0))\}d\lambda\right| \quad &\le\quad \|\mathrm{tr}\{(\overline{h}^i - \widetilde{h}^i)f_\theta\}\|_v\|\mathrm{tr}\{f_\theta^{-1}(E(I_n(\lambda,d_0))-f_\theta)\}\|_u + \|\mathrm{tr}\{(\overline{h}^i - \widetilde{h})f_\theta\}\|_1\\ &\le\quad (\frac{a_1\epsilon}{4})2^{-l}.\end{aligned}$$

Since

$$\begin{aligned}\|\mathrm{tr}\{(\overline{h}^i - \widetilde{h})f\}\|_1 \quad &\le\quad \|\mathrm{tr}\{(\overline{h}^i - \widetilde{h})f\}\|_v\\ &\le\quad C_1\max_{\beta_1\beta_2}\Big\{|\|[k^*(\overline{h}^i - h^0)k]_{\beta_1\beta_2}\|_v + |\|[k^*(\widetilde{h}^i - h^0)k]_{\beta_1\beta_2}\|_v\Big\}.\end{aligned}$$

Therefore, in view of Assumption (D3),

$$Pr\Big[\sup_{A(l)} T_n(h)/\{n^{-1/2} + |V(\theta)|\} > \epsilon\Big] \le m(\epsilon')\max_i P\Big[\sqrt{n}T_n(h^i) > \epsilon a_1\sqrt{n}2^{-(l+1)}\Big]. \tag{20}$$

For sufficiently large n,

$$\begin{aligned}Var(T_n(h^i)) \quad &\le\quad C_2\Big[\max_{\alpha_1\alpha_2}\|[(\overline{h^i} - h^0)f]_{\alpha_1\alpha_2}\|_2^2 + \max_{\beta_1\beta_2}\|[k^*(\overline{h^i} - h^0)k]_{\beta_1\beta_2}\|_2^2\Big]\\ &=\quad C_2Q(2^{-l}).\end{aligned}$$

Thus, following the Schwartz inequality, we can rewrite (20) as follows

$$\begin{aligned}m(\epsilon')\max_i P\Big[\sqrt{n}T_n(h^i) > \epsilon a_1\sqrt{n}2^{-(l+1)}\Big] \quad &\le\quad m(\epsilon')C_2Q(2^{-l})/\Big(\epsilon a_1\sqrt{n}2^{-(l+1)}\Big)^2\\ &=\quad m(\epsilon')C_2Q(2^{-l})4^{l+1}/n\epsilon^2.\end{aligned}$$

For $P\{\inf_{A(l)} T_n(h) > -\epsilon\}$, a similar limit can be used and we have

$$P\Big[\sup_{A(l)} |T_n(h)|/\{n^{-1/2} + |V(\theta)|\} > \epsilon\Big] \quad \leq \quad 8m(\epsilon^{'})C_2Q(2^{-l})4^l/n\epsilon^2.$$

Furthermore, we have

$$P\Big[\sup_{B(l_0)} |T_n(h)|/\{n^{-1/2} + |V(\theta)|\} > \epsilon\Big] \quad \leq \quad C_3Q(2^{-l_0})/\epsilon^2.$$

Set $l^{'}$ and $\epsilon^{'}$ such that, for $l \geq l^{'}$, $8m(\epsilon^{'})C_3Q(2^{-l})/\epsilon^2 < \epsilon$,

$$\begin{aligned} & P\Big[\sup_{B(l_0)} |T_n(h)|/\{n^{-1/2} + |V(\theta)|\} > \epsilon\Big] \\ \leq \quad & \Big(\sum_{l=0}^{l^{'}-1} + \sum_{l^{'}}^{l_0}\Big)P\Big[\sup_{A(l)} |T_n(h)|/\{n^{-1/2} + |V(\theta)|\} > \epsilon\Big] + P\Big[\sup_{B(l_0)} |T_n(h)|/\{n^{-1/2} + |V(\theta)|\} > \epsilon\Big] \\ \leq \quad & 8m(\epsilon^{'})C_3Q(1)(4^{l^{'}} - 1)/3n\epsilon^2 + \epsilon(4^{l_0+1} - 1)/3n + C_3Q(2^{l_0})/\epsilon^2. \end{aligned}$$

Since $l^{'}$ is independent of n, $8m(\epsilon^{'})C_3Q(1)(4^{l^{'}} - 1)/3n\epsilon^2 \to 0$ and $\epsilon(4^{l_0+1} - 1)/3n \to 0$ as $n \longrightarrow \infty$ moreover $C_3Q(2^{l_0})/\epsilon^2 < \epsilon$ and the result follows.

Proof of Theorem 2. Under Assumptions B and D, $\sqrt{n}S_n(\theta_0) \to^d N(0, U)$.

1) We consider that each element of cov($\sqrt{n}S_n(\theta_0)$) is defined by

$$K_n(\lambda_1\lambda_2\lambda_3\lambda_4) \quad = \quad \frac{1}{n}\prod_{j=1}^{4}\Big\{\frac{1}{\sqrt{2\pi}}\sum_{t=1}^{n}\exp(it\lambda_j)\Big\},$$

where $\varphi_n(\lambda_1) = \frac{1}{\sqrt{2\pi}}\sum_{t=1}^{n}\exp(it\lambda_1)$ and $K_n(\lambda_1\lambda_2\lambda_3\lambda_4) = \frac{1}{n}\varphi_n(\lambda_1)\varphi_n(\lambda_2)\varphi_n(\lambda_3)\varphi_n(\lambda_4)$.

As in (Hosoya, 1997), following $\{\lambda_1-\lambda_2+\lambda_3-\lambda_4 = 0\}$ as $n \to \infty$ and Assumption (D4), we have $\lim_{n\to\infty} \mathrm{cov}\{ \sqrt{n}S_n(\theta_0)\}$ $= U$ (see Hosoya, 1993).

2) If f_θ is square-integrable, Theorem 2 is satisfied (Hosoya & Taniguchi, 1982). So, if f_θ is non-square-integrable, we consider that

$$f_\theta(\lambda) \quad = \quad \Gamma(\exp(-i\lambda))\Gamma(\exp(-i\lambda))^*,$$

where $\Gamma(\exp(-i\lambda))$ is the boundary value of a $m \times m$-matrix-valued analytic function $\Gamma(z)$ in the unit disk (see Rozanov, 1967). Using this Γ, set $k^{'}(\lambda) = \Gamma^{-1}(\exp(-i\lambda))k(\lambda)$ and $h_j^{'}(\lambda) = \Gamma(\exp(-i\lambda))^*h_j(\lambda)\Gamma(\exp(-i\lambda))$. Then, we define the coefficients G_j' by $k'(\lambda) \equiv \sum_{j=0}^{\infty} G^{'}(j,\theta)\varepsilon(t-j)$ construct a new process $\{Z^{'}(t)\}$ by

$$Z^{'}(t) \quad = \quad \sum_{j=0}^{\infty} G^{'}(j,\theta)\varepsilon(t-j).$$

As in (Hosoya, 1997, Theorem 1.2)

$$\int_{-\pi}^{\pi}\mathrm{tr}\{h_j(\lambda)I_n(\lambda, d_0)\}d\lambda \quad \to \quad \int_{-\pi}^{\pi}\mathrm{tr}\{h_j(\lambda)f(\lambda)\}d\lambda$$

as $n \to \infty$ and

$$\mathrm{tr}\{h(\lambda)f(\lambda)h(\lambda)f(\lambda)\} = \mathrm{tr}\{h_j^{'}(\lambda)f^{'}(\lambda)h_j^{'}(\lambda)f^{'}(\lambda)\} = \mathrm{tr}\{h_j(\lambda)k(\lambda)^*h_j(\lambda)k(\lambda)\},$$

with $f^{'}(\lambda) = \Gamma^{-1}(\exp(-i\lambda))f(\lambda)\Gamma^{-1}(\exp(-i\lambda))^*$ and $k(\lambda)^*h_j(\lambda)k(\lambda) = k^{'}(\lambda)^*h_j^{'}(\lambda)k^{'}(\lambda)$, then

$$\lim_{n\to\infty} Var\Big[\sqrt{n}\int_{-\pi}^{\pi}\mathrm{tr}\{h_j(\lambda)I_n(\lambda, d_0)\}d\lambda\Big] = Var\Big[\sqrt{n}\int_{-\pi}^{\pi}\mathrm{tr}\{h_j^{'}(\lambda)I_n(\lambda, d_0^{'})\}d\lambda\Big].$$

Also the variance of the difference

$$\lim_{n\to\infty} n\Big\{Var\Big[\int_{-\pi}^{\pi}\mathrm{tr}\{h_j(\lambda)I_n(\lambda,d_0)\}d\lambda\Big]-Var\Big[\int_{-\pi}^{\pi}\mathrm{tr}\{h_j^{'}(\lambda)I_n(\lambda,d_0^{'})\}d\lambda\Big]\Big\}=0,$$

for all $j=1,\dots,p+q+1$, where

$$\begin{aligned}
&\lim_{n\to\infty} nVar\Big[\int_{-\pi}^{\pi}\mathrm{tr}\{h_j(\lambda)I_n(\lambda,d_0)\}d\lambda\Big]\\
= \;& 4\pi\int_{-\pi}^{\pi}\mathrm{tr}\{h_j(\lambda,\theta)f_j(\lambda)h_j(\lambda,\theta)f_j(\lambda)\}d\lambda+2\pi\sum_{\beta_1,\dots,\beta_4=1}^{m}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\Big[k^*(\lambda_1)h_j(\lambda_1,\theta_0)k(\lambda_1)\Big]_{\beta_1\beta_2}\\
&\times\Big[k^*(\lambda_2)h_j(\lambda_2,\theta_0)k(\lambda_2)\Big]_{\beta_3\beta_4}\times Q^{\varepsilon}_{\beta_1,\dots,\beta_4}(\lambda_1,\lambda_2,-\lambda_2)d\lambda_1 d\lambda_2\\
= \;& U\\
= \;& \lim_{n\to\infty} n\mathrm{cov}(S_n(\theta_0)).
\end{aligned}$$

And $\sqrt{n}\int_{-\pi}^{\pi}\mathrm{tr}\Big\{h_j^{'}(\lambda)[I_n(\lambda,d_0^{'})-E(I_n(\lambda,d_0^{'}))]\Big\}d\lambda$ is asymptotically normal if the pseudo spectral density of $\{Z^{'}(t)\}$ is square-integrable.

3) And then, from Lemma 1, we have the result following these relations $\sqrt{n}(\widetilde{\theta}_n-\theta_0)=W^{-1}\sqrt{n}S_n(\theta_0)+o_p(1)$, for $\sqrt{n}S_n(\theta_0)\to N(0,U)$ and $\sqrt{n}(\widetilde{\theta}_n-\theta_0)\to N\Big(0,W^{-1}U(W^*)^{-1}\Big)$ implies that $\sqrt{n}(\widetilde{\theta}_n-\theta_0)/W^{-1}\to N(0,U)$.

5. Simulations

In this section, we expose the Quasi-Maximum Likelihood Estimate (QMLE) and the Extended Multivariate Local Whittle Estimate (ExtMLWE). We consider the model ARFIMA$(0,d,0)$ for our simulation and we generate non-stationary multivariate ARFIMA processes by truncating the moving average representation in (4). We allow the fractional parameter of interest $(d_{1,0},d_{2,0})$ to belong to the set $\{(0.6,0.7),(1.1,1.2),(1.4,1.3),(1.6,1.7)\}$. The sample size $n\in\{100,200\}$. The bias and the root mean squared error (RMSE) are computed using 1000 replications. We obtain the following tables:

Table 1. Simulation results for bias and RMSE of QMLE and ExtMLWE, with $n=100$

	QMLE			ExtMLWE	
	$n=100$			$n=100$	
d_0	Bias	RMSE	d_0	Bias	RMSE
$d_{1,0}=0.6$	0.01	0.00636	$d_{1,0}=0.6$	-0.01	0.0068201
$d_{2,0}=0.7$	0.01	0.004459	$d_{2,0}=0.7$	-0.01	0.00577
$d_{1,0}=1.1$	0.01	0.00129	$d_{1,0}=1.1$	-0.011	0.00334
$d_{2,0}=1.2$	-0.0002	0.000227	$d_{2,0}=1.2$	-0.003	0.000933
$d_{1,0}=1.4$	0.0001	0.00179	$d_{1,0}=1.4$	0.01	0.001874
$d_{2,0}=1.3$	0.001	0.00029	$d_{2,0}=1.3$	-0.01	0.00072
$d_{1,0}=1.6$	0.001	0.000127	$d_{1,0}=1.6$	-0.001	0.000534
$d_{2,0}=1.7$	-0.001	0.000132	$d_{2,0}=1.7$	-0.01	0.000202

Table 2. Simulation results for bias and RMSE of QMLE and ExtMLWE, with $n = 200$

	QMLE			ExtMLWE	
	$n = 200$			$n = 200$	
d_0	Bias	RMSE	d_0	Bias	RMSE
$d_{1,0} = 0.6$	0.002	0.001425	$d_{1,0} = 0.6$	-0.011	0.001673
$d_{2,0} = 0.7$	0.001	0.00187	$d_{2,0} = 0.7$	0.001	0.00427
$d_{1,0} = 1.1$	0.001	0.00105	$d_{1,0} = 1.1$	0.002	0.001286
$d_{2,0} = 1.2$	0.0019	0.00157	$d_{2,0} = 1.2$	-0.001	0.0046486
$d_{1,0} = 1.4$	-0.001	0.000217	$d_{1,0} = 1.4$	-0.01	0.000466
$d_{2,0} = 1.3$	-0.001	0.00045	$d_{2,0} = 1.3$	-0.01	0.000615
$d_{1,0} = 1.6$	0.001	0.000898	$d_{1,0} = 1.6$	-0.001	0.001547
$d_{2,0} = 1.7$	-0.001	0.000863	$d_{2,0} = 1.7$	-0.002	0.001223

In general, according to Tables 1 and 2, the QMLE and the ExtMLWE present a good result for the bias. Thus, the QMLE improves the estimates because the bias and the RMSE are in general quite lower than that of the ExtMLWE.

Acknowledgements

We are grateful to the anonymous referees whose suggestions helped us to improve the quality of our paper.

References

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis*. Forcasting and Control, San Francisco, Holden Day.

Granger, C. W. J. (1980). Long-memory relationships and the aggregation of dynamic models. *Journal of Econometrics, 14*, 227-238. http://dx.doi.org/10.1016/0304-4076(80)90092-5

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-memory time series and fractional differencing. *Journal of Time Series Analysis, 1*, 15-29. http://dx.doi.org/10.1111/j.1467-9892.1980.tb00297.x

Hannan, E. J. (1970). *Multiple Time Series*. New York: Wiley. http://dx.doi.org/10.1002/9780470316429

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika, 68*, 165-176. http://dx.doi.org/10.1093/biomet/68.1.165

Hosoya, Y. (1989). The bracketing condition for limit theorems on stationary linear processes. *The Annals of Statistics, 17*, 401-418. http://dx.doi.org/10.1214/aos/1176347024

Hosoya, Y. (1993). A limit theory on stationary process with long-range and related statistical models. *Discussion paper* (Faculty of economics, Tohoku University, Sendal), 106.

Hosoya, Y. (1996). The quasi-likelihood approach to statistical inference on multiple times-series with long-range dependence. *Journal of Econometrics, 73*, 217-236. http://dx.doi.org/10.1016/0304-4076(95)01738-0

Hosoya, Y. (1997). A limit theory for long-range dependence and statistical inference on related models. *The Annals of Statistics, 25*(1), 105-137. http://dx.doi.org/10.1214/aos/1034276623

Hosoya, Y., & Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *The Annals of Statistics, 10*, 132-153. http://dx.doi.org/10.1214/aos/1176345696

Huber, P. (1967). The behavior of maximum-likelihood estimates under nonstandard conditions. *Procedings of the fifth Berkeley Symposium on Math. Statist. Probability, 1*, 221-231. Berkeley: Univ. California Press.

Nielsen, F. S. (2009). Local whittle estimation of multivariate fractionally integrated processes. *CREASTES Research Paper, 38*.

Odaki, M. (1993). On the invertibility of fractionally differenced ARIMA processes. *Biometrika, 80*, 703-709. http://dx.doi.org/10.1093/biomet/80.3.703

Phillips, P. C. B., & Shimotsu, K. (2004). Local Whittle Estimation in Nonstationary and Unit Root Cases. *Annals of Statistics, 32*, 656-692. http://dx.doi.org/10.1214/009053604000000139

Rozanov, Y. A. (1967). *Stationary Random Process*. Holden-Day, San Francisco.

Shao, X. F. (2009). *Nonstationarity-extended whittle estimation*. University of Illinois at Urbana-Champaign.

Shimotsu, K. (2007). Gaussian semiparametric estimation of multivariate fractionally integrated processes. *Journal of Econometrics, 137*, 277-310. http://dx.doi.org/10.1016/j.jeconom.2006.01.003

Sowell, F. (1987). *Maximum likelihood estimation of fractionally integrated time series models.* Discussion Paper 87-07, Department of economics, Duke University.

Sowell, F. (1989). *A decomposition of block Toeplitz matrices with applications to vector time series*. Discussion Paper, GSIA, Carnegie Mellon University.

Velasco, C. (1999a). Gaussian semiparametric estimation of non-stationary time series. *Annals of Statistics, 33*, 87-127.

A Marshall-Olkin Power Log-normal Distribution and Its Applications to Survival Data

Wenhao Gui[1]

[1] Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth MN, USA

Correspondence: Wenhao Gui, Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth MN 55812, USA. Tel: 1-218-726-7200. E-mail: wgui@d.umn.edu

Received: December 4, 2012 Accepted: December 23, 2012 Online Published: January 16, 2013

doi:10.5539/ijsp.v2n1p63 URL: http://dx.doi.org/10.5539/ijsp.v2n1p63

Abstract

In this paper, using Marshall-Olkin transformation, a new class of Extended Power Log-normal distribution which includes the Power Log-normal and Log-normal distributions as special cases is introduced. Its characterization and statistical properties are studied. A real survival dataset is analyzed and the results show that the proposed model is flexible and appropriate.

Keywords: power log-normal distribution, Marshall-Olkin transformation, survival analysis, maximum likelihood

1. Introduction

A Log-normal distribution is a well known continuous probability distribution of a random variable whose logarithm is normally distributed. In survival analysis, the lognormal distribution is extensively used in applications, for example, see Gupta et al. (1997), Royston (2001), Rutqvist (1985) and Johnson et al. (1996) etc. The density and cumulative distribution functions of a Log-normal random variable denoted by $X \sim LN(\mu, \sigma)$ are given by, for $-\infty < \mu < \infty, \sigma > 0, x > 0$,

$$f(x) = \frac{1}{x\sigma}\phi(\frac{\ln(x) - \mu}{\sigma}), \quad F(x) = \Phi(\frac{\ln(x) - \mu}{\sigma}), \tag{1}$$

where ϕ and Φ are the density and cumulative distribution functions of the standard normal distribution.

Nelson and Dognanksoy (1992) extended the Log-normal distribution and introduced the Power Log-normal distribution whose density and cumulative distribution functions are given by,

$$f(x) = \frac{p}{x\sigma}\phi(\frac{\mu - \ln(x)}{\sigma})[\Phi(\frac{\mu - \ln(x)}{\sigma})]^{p-1}, \quad F(x) = 1 - [\Phi(\frac{\mu - \ln(x)}{\sigma})]^{p}, \tag{2}$$

for $-\infty < \mu < \infty, \sigma > 0, p > 0, x > 0$. We denote it as $X \sim PLN(\mu, \sigma, p)$.

They fitted it to the life or strength data from specimens of various sizes. They presented that such a model arises when any specimen can be regarded as a series system of smaller portions, where portions of a certain size have a normal life (or strength) distribution. The statistical analysis can also be found in Nelson and Doganaksoy (1995). Szyszkowicz and Yanikomeroglu (2009) and Liu et al. (2008) proposed the use of power lognormal distributions to approximate lognormal sum distributions.

On the other hand, by adding a new parameter $\alpha > 0$ to an existing distribution, Marshall and Olkin (1997) proposed a new family of survival functions. The new parameter results in flexibility in the distribution. Let $\bar{F}(x) = 1 - F(x)$ be the survival function of a random variable X. Then

$$\bar{G}(x) = \frac{\alpha\bar{F}(x)}{1 - (1 - \alpha)\bar{F}(x)} \tag{3}$$

is a proper survival function. $\bar{G}(x)$ is called Marshall-Olkin family of distributions. If $\alpha = 1$, we have that $G = F$. The density function corresponding to (3) is given by

$$g(x) = \frac{\alpha f(x)}{[1 - (1 - \alpha)\bar{F}(x)]^2},$$

and the hazard rate function is given by

$$h(x) = \frac{h_F(x)}{1 - (1-\alpha)\bar{F}(x)},$$

where $h_F(x)$ is the hazard rate function of the original model with distribution F.

Using the Marshall-Olkin transformation (3), several researchers have studied various distribution extensions. Marshall and Olkin (1997) generalized the exponential and Weibull distributions. Alice and Jose (2003) introduced Marshall-Olkin extended semi Pareto model for Pareto type III and estabilished its geometric extreme stability. Semi-Weibull distribution and generalized Weibull distributions are discussed by Alice and Jose (2005). Ghitany et al. (2005) studied the Marshall-Olkin Weibull distribution, that can be obtained as a compound distribution mixing with exponential distribution, and applied it to model censored data. Marshall-Olkin Extended Lomax Distribution was introduced by Ghitany et al. (2007). Jose et al. (2010) investigated Marshall-Olkin q-Weibull distribution and its max-min processes. García et al. (2011) generalized the standard Log-normal distribution.

In this paper, we use the Marshall-Olkin transformation to define a new model, so-called the Marshall-Olkin Power Log-normal distribution (MPLN), which generalizes the Power Log-normal, the Log-normal model. We aim to reveal some statistical properties of the proposed model and apply it to survival analysis.

The rest of this article is organized as follows: in Section 2, we introduce the new defined distribution and investigate its basic properties, including the shape properties of its density function and the hazard rate function, stochastic orderings and representation, moments and measurements based on the moments. Section 3 discusses the estimation of parameters by the method of maximum likelihood. An application of the MPLN model to real survival data is illustrated in Section 4. Our work is concluded in Section 5.

2. Marshall-Olkin Power Log-normal Distribution and Its Properties

2.1 Density and Hazard Function

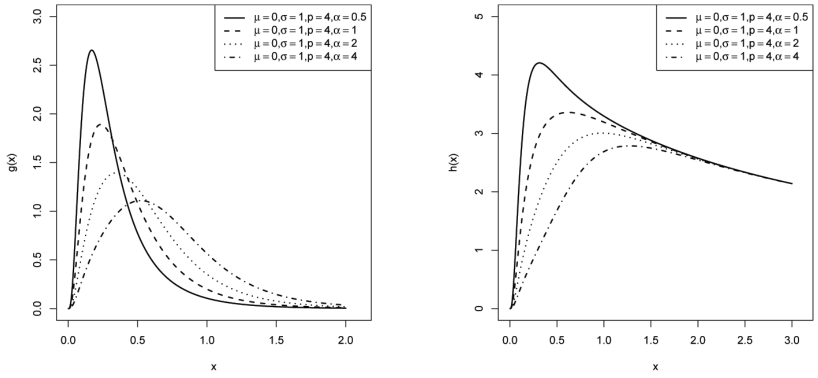
Let X follow the Power Log-normal distribution $PLN(\mu, \sigma, p)$, then its survival function is given by $\bar{F}(x) = 1 - F(x) = [\Phi(\frac{\mu-\ln(x)}{\sigma})]^p$. Substituting it in (3) we obtain a Marshall-Olkin Power Log-normal distribution denoted by $MPLN(\mu, \sigma, p, \alpha)$ with the following survival function

$$\bar{G}(x) = \frac{\alpha[\Phi(\frac{\mu-\ln(x)}{\sigma})]^p}{1 - (1-\alpha)[\Phi(\frac{\mu-\ln(x)}{\sigma})]^p}, \quad x > 0. \tag{4}$$

The corresponding density function is given by

$$g(x) = \frac{p\alpha[\Phi(\frac{\mu-\ln(x)}{\sigma})]^{p-1}\phi(\frac{\mu-\ln(x)}{\sigma})}{x\sigma\left((\alpha-1)[\Phi(\frac{\mu-\ln(x)}{\sigma})]^p + 1\right)^2}, \quad x > 0. \tag{5}$$

If $\alpha = 1$, we obtain the Power Log-normal distribution with parameter μ, σ, p. Furthermore, if $p = 1$, it reduces to the Log-normal distribution. This distribution contains the Power Log-normal distribution and Log-normal distribution as particular cases.



(a) Density function

(b) Hazard function

Figure 1. Plots of Marshall-Olkin power log-normal density and hazard function for some parameter values

Figure 1(a) shows some density functions of the $MPLN(\mu, \sigma, p, \alpha)$ distribution with various parameters. It indicates that the value of α has a subtantial effect on the tail of the density function.

The hazard rate function of the $MPLN(\mu, \sigma, p, \alpha)$ distribution is given by

$$h(x) = \frac{g(x)}{\bar{G}(x)} = \frac{p\phi(\frac{\mu-\ln(x)}{\sigma})}{x\sigma\Phi(\frac{\mu-\ln(x)}{\sigma})[(\alpha-1)[\Phi(\frac{\mu-\ln(x)}{\sigma})]^p+1]}, \quad x > 0. \tag{6}$$

Figure 1(b) shows some shapes of the $MPLN(\mu, \sigma, p, \alpha)$ hazard function with various parameters.

2.2 Stochastic Orderings

In statistics, a stochastic order measures the concept of one random variable being "larger" than another. It is an important tool to judge the comparative behavior. Here are some basic definitions.

A random variable X is less than Y in the ususal stochastic order (denoted by $X \prec_{st} Y$) if $F_X(x) \geq F_Y(x)$ for all real x. X is less than Y in the hazard rate order (denoted by $X \prec_{hr} Y$) if $h_X(x) \geq h_Y(x)$, for all $x > 0$. X is less than Y in the likelihood ratio order (denoted by $X \prec_{lr} Y$) if $f_Y(x)/f_X(x)$ increases in $x > 0$. It is well known that $X \prec_{lr} Y \Rightarrow X \prec_{hr} Y \Rightarrow X \prec_{st} Y$, see Ramesh and Kirmani (1987).

Proposition 1 *If $X \sim MPLN(\mu, \sigma, p, \alpha_1)$ and $Y \sim MPLN(\mu, \sigma, p, \alpha_2)$, and $\alpha_1 < \alpha_2$, then $X \prec_{lr} Y$, $X \prec_{hr} Y$ and $X \prec_{st} Y$.*

Proof. The density ratio is given by

$$U(x) = \frac{f_X(x)}{f_Y(x)} = \frac{\alpha_1[1-(1-\alpha_2)\,\Phi^p(\frac{\mu-\ln(x)}{\sigma})]^2}{\alpha_2[1-(1-\alpha_1)\,\Phi^p(\frac{\mu-\ln(x)}{\sigma})]^2}.$$

Taking the derivative with respect to x,

$$U'(x) = \frac{2p\alpha_1(\alpha_1-\alpha_2)\Phi^{p-1}(\frac{\mu-\ln(x)}{\sigma})[(\alpha_2-1)\Phi^p(\frac{\mu-\ln(x)}{\sigma})+1]\phi(\frac{\mu-\ln(x)}{\sigma})}{x\sigma\alpha_2[(\alpha_1-1)\Phi^p(\frac{\mu-\ln(x)}{\sigma})+1]^3}.$$

If $\alpha_1 < \alpha_2$, $U'(x) < 0$, $U(x)$ is a decreasing function of x. The results follow.

2.3 Stochastic Representation

Let $\bar{G}_0(x|\lambda), -\infty < x < \infty, -\infty < \lambda < \infty$, be the conditional survival function of a continuous random variable X given a continuous random variable λ. Let Λ be a random variable with probability density function $m(\lambda)$. Then the distribution with survival function

$$\bar{G}(x) = \int_{-\infty}^{\infty} \bar{G}_0(x|\lambda)m(\lambda)d\lambda, \quad -\infty < x < \infty,$$

is called a compounding distribution with mixing density $m(\lambda)$. Compounding distribution provides a useful way to obtain new class of distributions in terms of existing ones. The following result shows that the $MPLN(\mu, \sigma, p, \alpha)$ distribution can be expressed as a compound distribution.

Proposition 2 *Suppose that the conditional survival function of a continuous random variable X given $\Lambda = \lambda$ is given by*

$$\bar{G}_0(x|\lambda) = \exp\left[-\lambda\Phi^{-p}(\frac{\mu-\ln(x)}{\sigma}) + \lambda\right], \quad x > 0. \tag{7}$$

Let Λ have an exponential distribution with density function

$$m(\lambda) = \alpha e^{-\alpha\lambda}, \quad \alpha > 0, \lambda > 0.$$

Then the random variable X has the $MPLN(\mu, \sigma, p, \alpha)$ distribution.

Proof. For $x > 0$, the survival function of X is given by

$$\bar{G}(x) = \int_0^{\infty} \bar{G}_0(x|\lambda)m(\lambda)d\lambda = \alpha\int_0^{\infty} e^{-\lambda\Phi^{-p}(\frac{\mu-\ln(x)}{\sigma})+\lambda}e^{-\alpha\lambda}d\lambda = \frac{\alpha[\Phi(\frac{\mu-\ln(x)}{\sigma})]^p}{1-(1-\alpha)[\Phi(\frac{\mu-\ln(x)}{\sigma})]^p},$$

which is the survival function of the $MPLN(\mu, \sigma, p, \alpha)$ distribution.

For $\lambda > 0$, $\bar{G}_0(x|\lambda)$ defines a class of non-standard distributions. Compounding a distribution belonging to this class with an exponential distribution for λ leads to a certain $MPLN(\mu, \sigma, p, \alpha)$ distribution. Next we will present another stochastic representation of the $MPLN(\mu, \sigma, p, \alpha)$ distribution.

Proposition 3 *Let $\{X_i, i \geq 1\}$ be a sequence of i.i.d. random variables with a Power Log-normal distribution $PLN(\mu, \sigma, p)$. Let N be a geometric random variable with parameter $0 < \alpha < 1$ such that $P(N = n) = \alpha(1 - \alpha)^{n-1}, n = 1, 2, \ldots,$ which is independent of $\{X_i, i \geq 1\}$. Then,*

(1) $\min(X_1, \ldots, X_N)$ has a Marshall-Olkin Power Log-normal distribution $MPLN(\mu, \sigma, p, \alpha)$.

(2) $\max(X_1, \ldots, X_N)$ has a Marshall-Olkin Power Log-normal distribution $MPLN(\mu, \sigma, p, 1/\alpha)$.

Proof. The survival function of $\min(X_1, \ldots, X_N)$ is

$$\begin{aligned} P(\min(X_1, \ldots, X_N) > x) &= \sum_{n=1}^{\infty} P(X_1 > x, \ldots, X_n > x)P(N = n) \\ &= \sum_{n=1}^{\infty} [\bar{F}(x)]^n \alpha(1 - \alpha)^{n-1} \\ &= \frac{\alpha \bar{F}(x)}{1 - (1 - \alpha)\bar{F}(x)}, \end{aligned}$$

which is the survival function of the Marshall-Olkin Power Log-normal distribution $MPLN(\mu, \sigma, p, \alpha)$.

The survival function of $\max(X_1, \ldots, X_N)$ is

$$\begin{aligned} P(\max(X_1, \ldots, X_N) > x) &= 1 - P(\max(X_1, \ldots, X_N) \leq x) \\ &= 1 - \sum_{n=1}^{\infty} P(X_1 \leq x, \ldots, X_n \leq x)P(N = n) \\ &= 1 - \sum_{n=1}^{\infty} [F(x)]^n \alpha(1 - \alpha)^{n-1} \\ &= \frac{\frac{1}{\alpha}\bar{F}(x)}{1 - (1 - \frac{1}{\alpha})\bar{F}(x)}, \end{aligned}$$

which is the survival function of the Marshall-Olkin Log-Logistic distribution $MPLN(\mu, \sigma, p, 1/\alpha)$.

2.4 Moments and Quantiles

Proposition 4 *Let $X \sim MPLN(\mu, \sigma, p, \alpha)$, for $k = 1, 2, \ldots,$ Then the kth non-central moment is given by*

$$\mu_k = E(X^k) = p\alpha e^{k\mu} \int_0^1 e^{-k\sigma\Phi^{-1}(h)} \frac{h^{p-1}}{[(\alpha - 1)h^p + 1]^2} dh, \tag{8}$$

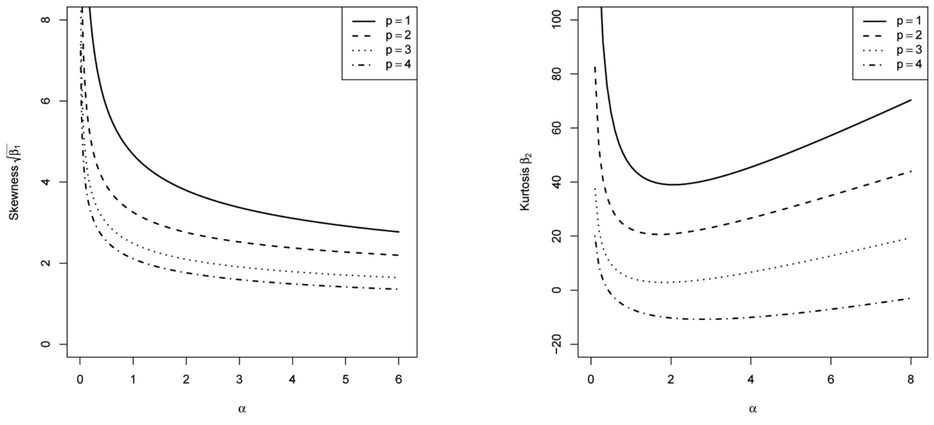
where Φ^{-1} is the inverse (quantile function) of the normal cumulative distribution function.

Proof. By definition of the moment,

$$\begin{aligned} \mu_k &= \int_0^{\infty} x^k g(x) dx \\ &= \int_0^{\infty} x^k \frac{p\alpha[\Phi(\frac{\mu - \ln(x)}{\sigma})]^{p-1} \phi(\frac{\mu - \ln(x)}{\sigma})}{x\sigma\left((\alpha - 1)[\Phi(\frac{\mu - \ln(x)}{\sigma})]^p + 1\right)^2} dx \quad let \quad t = (\mu - \ln(x))/\sigma \\ &= \int_{-\infty}^{\infty} e^{k(\mu - \sigma t)} \frac{p\alpha[\Phi(t)]^{p-1}\phi(t)}{((\alpha - 1)[\Phi(t)]^p + 1)^2} dt \quad let \quad h = \Phi(t) \\ &= p\alpha e^{k\mu} \int_0^1 e^{-k\sigma\Phi^{-1}(h)} \frac{h^{p-1}}{[(\alpha - 1)h^p + 1]^2} dh. \end{aligned}$$

The above expression seems to have no compact form. We can compute it with the help of computer. For the standardized skewness coefficient $\sqrt{\beta_1} = \frac{\mu_3 - 3\mu_1\mu_2 + 2\mu_1^3}{(\mu_2 - \mu_1^2)^{3/2}}$ and kurtosis coefficient $\beta_2 = \frac{\mu_4 - 4\mu_1\mu_3 + 6\mu_1^2\mu_2 - 3\mu_1^4}{(\mu_2 - \mu_1^2)^2}$, where

$\mu_1, \mu_2, \mu_3, \mu_4$ are the moments given in (8), Figure 2 shows the skewness and kurtosis coefficients for the Marshall-Olkin Power Log-normal $MPLN(\mu = 0, \sigma = 1, p, \alpha)$ model.



(a) Skewness coefficient

(b) Kurtosis coefficient

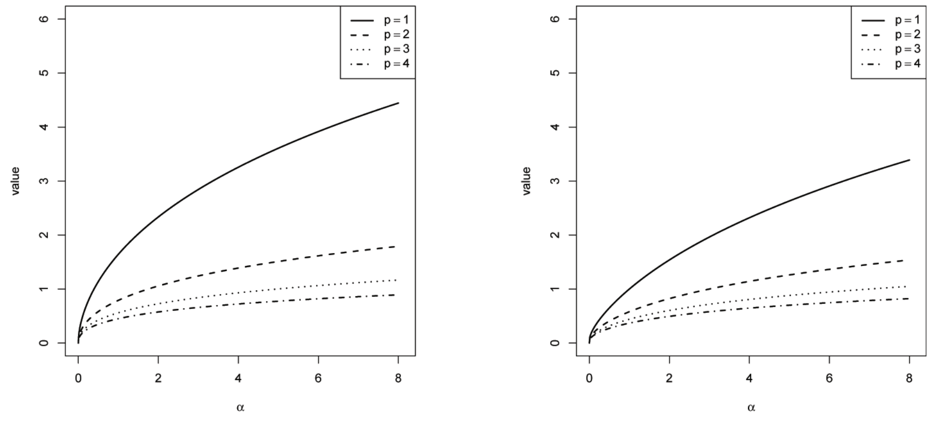
Figure 2. The plots for the skewness $\sqrt{\beta_1}$ and kurtosis coefficient β_2

The qth quantile $x_q = G^{-1}(q)$ of the $MPLN(\mu, \sigma, p, \alpha)$ distribution is given by

$$x_q = e^{\mu-\sigma\Phi^{-1}[(\frac{1-q}{\alpha q-q+1})^{\frac{1}{p}}]}, \quad 0 \leq q \leq 1. \tag{9}$$

where $G^{-1}(\cdot)$ is the inverse of distribution function. In particular, the median of the $MPLN(\mu, \sigma, p, \alpha)$ distribution is given by $median(X) = e^{\mu-\sigma\Phi^{-1}[(\frac{1}{\alpha+1})^{\frac{1}{p}}]}$.

Figure 3 displays the measures of central tendency (mean, median) of the $MPLN(\mu = 0, \sigma = 1, p, \alpha)$ distribution. From the figure, it is found that, as expected, the mean is larger than the median. The distribution has a right tail.



(a) Mean value

(b) Median value

Figure 3. Plots of mean and median of the $MPLN(\mu = 0, \sigma = 1, p, \alpha)$ distribution

3. Maximum Likelihood Estimation

In this section, we consider the maximum likelihood estimation about the parameters (μ, σ, p, α) of the Marshall-Olkin Power Log-normal model. Suppose $X_1, X_2, \ldots, X_n$ is a random sample of size n from the Marshall-Olkin Power Log-normal distribution $MPLN(\mu, \sigma, p, \alpha)$. Then the likelihood function is given by

$$\prod_{i=1}^{n} g_X(x_i) = \prod_{i=1}^{n} \frac{p\alpha[\Phi(\frac{\mu-\ln(x_i)}{\sigma})]^{p-1}\phi(\frac{\mu-\ln(x_i)}{\sigma})}{x_i\sigma\left((\alpha-1)[\Phi(\frac{\mu-\ln(x_i)}{\sigma})]^p+1\right)^2}, \tag{10}$$

and the log-likelihood function is given by

$$\begin{aligned} l &= \ln(\prod_{i=1}^{n} g_X(x_i)) \\ &= n\ln(p) + n\ln(\alpha) - n\ln(\sigma) + (p-1)\sum_{i=1}^{n}\ln(\Phi(\frac{\mu-\ln(x_i)}{\sigma})) + \sum_{i=1}^{n}\ln(\phi(\frac{\mu-\ln(x_i)}{\sigma})) \\ &\quad - \sum_{i=1}^{n}\ln(x_i) - 2\sum_{i=1}^{n}\ln[(\alpha-1)\Phi^p(\frac{\mu-\ln(x_i)}{\sigma}) + 1]. \end{aligned} \tag{11}$$

The estimates of the parameters maximize the likelihood function. Taking the partial derivatives of the log-likelihood function with respect to μ, σ, p, α respectively and equalizing the obtained expressions to zero yield to likelihood equations.

$$\frac{\partial l}{\partial \mu} = 0, \quad \frac{\partial l}{\partial \sigma} = 0, \quad \frac{\partial l}{\partial p} = 0, \quad \frac{\partial l}{\partial \alpha} = 0.$$

However, the equations do not lead to explicit analytical solutions for the parameters. Thus, the estimates must be obtained by means of numerical procedures such as Newton-Raphson method. The program R provides the nonlinear optimization function *optim* for solving such problems.

It is known that under some regular conditions, as the sample size increases, the distribution of the MLE tends to a multivariate normal distribution with mean $\theta = (\mu, \sigma, p, \alpha)^T$ and covariance matrix equal to the inverse of the Fisher information matrix $I^{-1}(\theta)$, see Cox and Hinkley (1979). The score vector and Hessian matrix are given in the Appendix. The multivariate normal distribution can be used to construct approximate confidence intervals for the parameters.

The likelihood ratio test can be used to test if the fit using MPLN model is statistically better than a fit using the PLN model. That is, we can test the hull hypothesis H_0: $\alpha = 1$ against H_1: $\alpha \neq 1$. When H_0 is true, the likelihood ratio statistic $d = 2[l(\hat{\mu}, \hat{\sigma}, \hat{p}, \hat{\alpha}) - l(\hat{\mu}, \hat{\sigma}, \hat{p}, 1)]$ has approximately a chi-square distribution with 1 degree of freedom, see Neyman and Pearson (1928) and Wilks (1938).

4. Application

In this section, we consider a real data set to illustrate the proposed model. The data taken from Davis (1952) are the number of miles to first and succeeding major motor failures of 191 buses operated by a large city bus company. The data is shown in Table 1.

Table 1. Initial bus motor failures

Distance interval(1000 miles)	Observed number of failures
Less than 20	6
20-40	11
40-60	16
60-80	25
80-100	34
100-120	46
120-140	33
140-160	16
160-180	2
180-up	2

We fit the data set with the Log-normal(LN), the Power Log-normal (PLN) and the Marshall-Olkin Power Log-normal(MPLN) distributions, respectively, using maximum likelihood method. The results are reported in Table 2. The usual Akaike information criterion (AIC) introduced by Akaike (1973) and Bayesian information criterion (BIC) proposed by Schwarz (1978) to measure of the goodness of fit are also computed. $AIC = 2k - 2\ln(L)$ and $BIC = k\ln(n) - 2\ln(L)$. where k is the number of parameters in the distribution and L is the maximized value of the likelihood function.

The results show that MPLN model fits best. Figure 4 displays the histogram and fitted models using the MLE estimates.

Table 2. Maximum likelihood parameter estimates (with standard deviation) of the LN, PLN and MPLN models for the initial bus motor failure data

Model	$\hat{\mu}$	$\hat{\sigma}$	$\hat{p}$	$\hat{\alpha}$	loglik	AIC	BIC
LN	4.454	0.566	–	–	−1013.410	2030.820	2037.325
	(0.141)	(0.290)					
PLN	12.312	1.729	19.714	–	−971.501	1949.002	1958.759
	(0.437)	(1.002)	(1.208)				
MPLN	15.435	2.671	10.558	12.454	−960.498	1928.996	1942.005
	(0.529)	(0.408)	(1.604)	(1.109)			

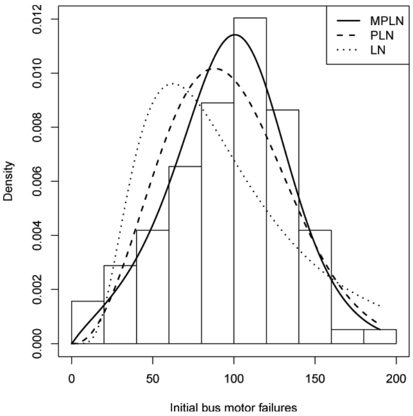


Figure 4. Histogram and fitted curves for the initial bus motor failure data

5. Conclusions

In this paper, the Power Log-normal distribution is generalized by adding an extra parameter. It is achieved by using the well known Marshall-Olkin transformation. The new model, named Marshall-Olkin Power Log-normal distribution, includes the Power Log-normal and Log-normal distributions as special cases.

Its detailed characterization and statistical properties such as stochastic orderings, stochastic representation, the moments and measures based on the moments, are presented. The estimation of parameters is approached by the method of maximum likelihood and the Hessian matrix is derived. A real survival dataset is analyzed and the results show that the proposed model is flexible and appropriate.

6. Appendix: Score Vector and Hessian Matrix

Suppose $x_1, x_2, ..., x_n$ is a random sample from the $MPLN(\mu, \sigma, p, \alpha)$ distribution, then the log-likelihood function is given by (11). The elements of the score vector are obtained by differentiation

$$\begin{aligned}
l_\mu &= -\sum_{i=1}^{n} \frac{2p(\alpha-1)\phi(\frac{\mu-\ln(x_i)}{\sigma})\Phi^{p-1}(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma[(\alpha-1)\Phi^{p}(\frac{\mu-\ln(x_i)}{\sigma})+1]} + \sum_{i=1}^{n} \frac{\phi'(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma\phi(\frac{\mu-\ln(x_i)}{\sigma})} + \sum_{i=1}^{n} \frac{(p-1)\phi(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma\Phi(\frac{\mu-\ln(x_i)}{\sigma})}, \\
l_\sigma &= \sum_{i=1}^{n} \frac{2p(\alpha-1)(\mu-\ln(x_i))\phi(\frac{\mu-\ln(x_i)}{\sigma})\Phi^{p-1}(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma^2[(\alpha-1)\Phi^{p}(\frac{\mu-\ln(x_i)}{\sigma})+1]} - \sum_{i=1}^{n} \frac{(\mu-\ln(x_i))\phi(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma^2\phi(\frac{\mu-\ln(x_i)}{\sigma})} - \frac{n}{\sigma} \\
&\quad -\sum_{i=1}^{n} \frac{(p-1)(\mu-\ln(x_i))\phi(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma^2\Phi(\frac{\mu-\ln(x_i)}{\sigma})}, \\
l_p &= -\sum_{i=1}^{n} \frac{2(\alpha-1)\ln[\Phi(\frac{\mu-\ln(x_i)}{\sigma})]\Phi^{p}(\frac{\mu-\ln(x_i)}{\sigma})}{(\alpha-1)\Phi^{p}(\frac{\mu-\ln(x_i)}{\sigma})+1} + \sum_{i=1}^{n} \ln[\Phi(\frac{\mu-\ln(x_i)}{\sigma})] + \frac{n}{p}, \\
l_\alpha &= \frac{n}{\alpha} - \sum_{i=1}^{n} \frac{2\Phi^{p}(\frac{\mu-\ln(x_i)}{\sigma})}{(\alpha-1)\Phi^{p}(\frac{\mu-\ln(x_i)}{\sigma})+1}.
\end{aligned}$$

The Hessian matrix, second partial derivatives of the log-likelihood, is given by

$$H(\theta)=\begin{pmatrix} l_{\mu\mu} & l_{\mu\sigma} & l_{\mu p} & l_{\mu\alpha}\\ l_{\sigma\mu} & l_{\sigma\sigma} & l_{\sigma p} & l_{\sigma\alpha}\\ l_{p\mu} & l_{p\sigma} & l_{pp} & l_{p\alpha}\\ l_{\alpha\mu} & l_{\alpha\sigma} & l_{\alpha p} & l_{\alpha\alpha}\end{pmatrix}$$

where

$$\begin{aligned}
l_{\mu\mu} &= -\sum_{i=1}^{n}\frac{2(p-1)p(\alpha-1)\phi(\frac{\mu-\ln(x_i)}{\sigma})^2\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^{p-2}}{\sigma^2\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p+1\right)}-\sum_{i=1}^{n}\frac{2p(\alpha-1)\phi'(\frac{\mu-\ln(x_i)}{\sigma})\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^{p-1}}{\sigma^2\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p+1\right)}\\
&\quad+\sum_{i=1}^{n}\frac{2p^2(\alpha-1)^2\phi(\frac{\mu-\ln(x_i)}{\sigma})^2\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^{2p-2}}{\sigma^2\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p+1\right)^2}+\sum_{i=1}^{n}\frac{\phi''(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma^2\phi(\frac{\mu-\ln(x_i)}{\sigma})}-\sum_{i=1}^{n}\frac{[\phi'(\frac{\mu-\ln(x_i)}{\sigma})]^2}{\sigma^2[\phi(\frac{\mu-\ln(x_i)}{\sigma})]^2}\\
&\quad+\sum_{i=1}^{n}\frac{(p-1)\phi'(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma^2\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}-\sum_{i=1}^{n}\frac{(p-1)\phi(\frac{\mu-\ln(x_i)}{\sigma})^2}{\sigma^2\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^2},\\
l_{\mu\sigma} &= l_{\sigma\mu}=\sum_{i=1}^{n}\frac{2(p-1)p(\alpha-1)(\mu-\ln(x_i))\phi(\frac{\mu-\ln(x_i)}{\sigma})^2\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-2}}{\sigma^3\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p+1\right)}\\
&\quad+\sum_{i=1}^{n}\frac{2p(\alpha-1)\phi(\frac{\mu-\ln(x_i)}{\sigma})\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-1}}{\sigma^2\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p+1\right)}+\sum_{i=1}^{n}\frac{2p(\alpha-1)(\mu-\ln(x_i))\phi'\left(\frac{\mu-\ln(x_i)}{\sigma}\right)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-1}}{\sigma^3\left((\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^p+1\right)}\\
&\quad-\sum_{i=1}^{n}\frac{2p^2(\alpha-1)^2(\mu-\ln(x_i))\phi(\frac{\mu-\ln(x_i)}{\sigma})^2\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{2p-2}}{\sigma^3\left((\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^p+1\right)^2}+\sum_{i=1}^{n}\frac{(\mu-\ln(x_i))[\phi'\left(\frac{\mu-\ln(x_i)}{\sigma}\right)]^2}{\sigma^3[\phi(\frac{\mu-\ln(x_i)}{\sigma})]^2}\\
&\quad-\sum_{i=1}^{n}\frac{(\mu-\ln(x_i))\phi''\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}{\sigma^3\phi(\frac{\mu-\ln(x_i)}{\sigma})}-\sum_{i=1}^{n}\frac{(p-1)\phi(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma^2\Phi(\frac{\mu-\ln(x_i)}{\sigma})}-\sum_{i=1}^{n}\frac{(p-1)(\mu-\ln(x_i))\phi'\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}{\sigma^3\Phi(\frac{\mu-\ln(x_i)}{\sigma})}\\
&\quad+\sum_{i=1}^{n}\frac{(p-1)\,(\mu-\ln(x_i))\,\phi(\frac{\mu-\ln(x_i)}{\sigma})^2}{\sigma^3\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^2}-\sum_{i=1}^{n}\frac{\phi'(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma^2\phi(\frac{\mu-\ln(x_i)}{\sigma})},\\
l_{\mu p} &= l_{p\mu}=-\sum_{i=1}^{n}\frac{2p(\alpha-1)\ln\left(\Phi(\frac{\mu-\ln(x_i)}{\sigma})\right)\phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-1}}{\sigma\left((\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^p+1\right)}-\sum_{i=1}^{n}\frac{2(\alpha-1)\phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-1}}{\sigma\left((\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^p+1\right)}\\
&\quad+\sum_{i=1}^{n}\frac{2p(\alpha-1)^2\ln\left(\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)\right)\phi(\frac{\mu-\ln(x_i)}{\sigma})\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^{2p-1}}{\sigma\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p+1\right)^2}+\sum_{i=1}^{n}\frac{\phi(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)},\\
l_{\mu\alpha} &= l_{\alpha\mu}=\sum_{i=1}^{n}\frac{2p(\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{2p-1}\phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}{\sigma\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p+1\right)^2}-\sum_{i=1}^{n}\frac{2p\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-1}\phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}{\sigma\left((\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^p+1\right)},\\
l_{\sigma\sigma} &= -\sum_{i=1}^{n}\frac{2(p-1)p(\alpha-1)(\mu-\ln(x_i))^2\phi(\frac{\mu-\ln(x_i)}{\sigma})^2\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-2}}{\sigma^4\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p+1\right)}\\
&\quad-\sum_{i=1}^{n}\frac{4p(\alpha-1)(\mu-\ln(x_i))\phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-1}}{\sigma^3\left((\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^p+1\right)}-\sum_{i=1}^{n}\frac{2p(\alpha-1)(\mu-\ln(x_i))^2\phi'(\frac{\mu-\ln(x_i)}{\sigma})\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^{p-1}}{\sigma^4\left((\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^p+1\right)}\\
&\quad+\sum_{i=1}^{n}\frac{2p^2(\alpha-1)^2(\mu-\ln(x_i))^2\phi(\frac{\mu-\ln(x_i)}{\sigma})^2\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{2p-2}}{\sigma^4\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p+1\right)^2}+\sum_{i=1}^{n}\frac{2(\mu-\ln(x_i))\phi'\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}{\sigma^3\phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}\\
&\quad-\sum_{i=1}^{n}\frac{(\mu-\ln(x_i))^2[\phi'\left(\frac{\mu-\ln(x_i)}{\sigma}\right)]^2}{\sigma^4[\phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)]^2}+\sum_{i=1}^{n}\frac{(\mu-\ln(x_i))^2\phi''\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}{\sigma^4\phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}+\frac{n}{\sigma^2}+\sum_{i=1}^{n}\frac{2(p-1)(\mu-\ln(x_i))\Phi'(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma^3\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}
\end{aligned}$$

$$\begin{aligned}
& + \sum_{i=1}^{n} \frac{(p-1)(\mu-\ln(x_i))^2 \phi'(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma^4 \Phi(\frac{\mu-\ln(x_i)}{\sigma})} - \sum_{i=1}^{n} \frac{(p-1)(\mu-\ln(x_i))^2 \Phi'(\frac{\mu-\ln(x_i)}{\sigma})^2}{\sigma^4 \Phi(\frac{\mu-\ln(x_i)}{\sigma})^2}, \\
l_{\sigma p} \quad = \quad & l_{p\sigma} = \sum_{i=1}^{n} \frac{2(\alpha-1)(\mu-\ln(x_i))\phi(\frac{\mu-\ln(x_i)}{\sigma})\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-1}}{\sigma^2 \left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p + 1\right)} \\
& + \sum_{i=1}^{n} \frac{2p(\alpha-1)(\mu-\ln(x_i)) \ln\left(\Phi(\frac{\mu-\ln(x_i)}{\sigma})\right) \phi(\frac{\mu-\ln(x_i)}{\sigma})\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-1}}{\sigma^2 \left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p + 1\right)} \\
& - \sum_{i=1}^{n} \frac{2p(\alpha-1)^2(\mu-\ln(x_i)) \ln\left(\Phi(\frac{\mu-\ln(x_i)}{\sigma})\right) \phi(\frac{\mu-\ln(x_i)}{\sigma})\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{2p-1}}{\sigma^2 \left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p + 1\right)^2} - \sum_{i=1}^{n} \frac{(\mu-\ln(x_i))\phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}{\sigma^2 \Phi(\frac{\mu-\ln(x_i)}{\sigma})}, \\
l_{\sigma\alpha} \quad = \quad & l_{\alpha\sigma} = \sum_{i=1}^{n} \frac{2p(\mu-\ln(x_i))\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{p-1}\phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)}{\sigma^2 \left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p + 1\right)} - \sum_{i=1}^{n} \frac{2p(\alpha-1)(\mu-\ln(x_i))\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^{2p-1}\phi(\frac{\mu-\ln(x_i)}{\sigma})}{\sigma^2 \left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p + 1\right)^2}, \\
l_{pp} \quad = \quad & -\sum_{i=1}^{n} \frac{2(\alpha-1)\ln^2\left(\Phi(\frac{\mu-\ln(x_i)}{\sigma})\right)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p}{(\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^p + 1} + \sum_{i=1}^{n} \frac{2(\alpha-1)^2\ln^2\left(\Phi(\frac{\mu-\ln(x_i)}{\sigma})\right)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^{2p}}{\left((\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^p + 1\right)^2} - \frac{n}{p^2}, \\
l_{p\alpha} \quad = \quad & l_{\alpha p} = \sum_{i=1}^{n} \frac{2(\alpha-1)\ln\left(\Phi(\frac{\mu-\ln(x_i)}{\sigma})\right)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^{2p}}{\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p + 1\right)^2} - \sum_{i=1}^{n} \frac{2\ln\left(\Phi(\frac{\mu-\ln(x_i)}{\sigma})\right)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p}{(\alpha-1)\Phi(\frac{\mu-\ln(x_i)}{\sigma})^p + 1}, \\
l_{\alpha\alpha} \quad = \quad & \sum_{i=1}^{n} \frac{2\Phi(\frac{\mu-\ln(x_i)}{\sigma})^{2p}}{\left((\alpha-1)\Phi\left(\frac{\mu-\ln(x_i)}{\sigma}\right)^p + 1\right)^2} - \frac{n}{\alpha^2}.
\end{aligned}$$

The Fisher information matrix $I(\theta) = -E(H(\theta))$.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory* (2 nd.), *Tsahkadsor, Armenian SSR*, 267-281. http://dx.doi.org/10.1007/978-1-4612-0919-5_38

Alice, T., & Jose, K. (2003). Marshall-olkin pareto processes. *Far East Journal of Theoretical Statistics, 9*(2), 117-132.

Alice, T., & Jose, K. (2005). Marshall-olkin semi-weibull minification processes. *Recent Advances in Statistical Theory and Applications, I*, 6-17.

Cox, D., & Hinkley, D. (1979). *Theoretical statistics*. Chapman & Hall/CRC.

Davis, D. (1952). An analysis of some failure data. *Journal of the American Statistical Association, 47*(258), 113-150. http://dx.doi.org/10.2307/2280740

García, V., Gómez-Déniz, E., & Vázquez-Polo, F. (2011). A generalization of the log–normal distribution and its applications. *4th Workshop on Risk Management and Insurance*.

Ghitany, M., Al-Awadhi, F., & Alkhalfan, L. (2007). Marshall-olkin extended lomax distribution and its application to censored data. *Communications in Statistics Theory and Methods, 36*(10), 1855-1866. http://dx.doi.org/10.1080/03610920601126571

Ghitany, M., Al-Hussaini, E., & Al-Jarallah, R. (2005). Marshall-olkin extended weibull distribution and its application to censored data. *Journal of Applied Statistics, 32*(10), 1025-1034. http://dx.doi.org/10.1080/02664760500165008

Gupta, R., Kannan, N., & Raychaudhuri, A. (1997). Analysis of lognormal survival data. *Mathematical biosciences, 139*(2), 103-115. http://dx.doi.org/10.1016/S0025-5564(96)00133-2

Johnson, W., Lee, J., Zellner, A., & Johnson, W. (1996). Predictive influence in the lognormal survival model. *Prediction and Modeling in Statistics and Econometrics: Essays in Honor of Seymour Geisser*.

Jose, K., Naik, S., & Ristić, M. (2010). Marshall-olkin q-weibull distribution and max-min processes. *Statistical papers, 51*(4), 837-851. http://dx.doi.org/10.1007/s00362-008-0173-9

Liu, Z., Almhana, J., & McGorman, R. (2008). Approximating lognormal sum distributions with power lognormal distributions. *Vehicular Technology, IEEE Transactions on, 57*(4), 2611-2617. http://dx.doi.org/10.1109/TVT.2007.912338

Marshall, A., & Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and weibull families. *Biometrika, 84*(3), 641-652. http://dx.doi.org/10.1093/biomet/84.3.641

Nelson, W., & Doganaksoy, N. (1995). Statistical analysis of life or strength data from specimens of various sizes using the power-(log) normal model. *Recent Advances in Life-Testing and Reliability*, CRC Press, 377-408.

Nelson, W., & Dognanksoy, N. (1992). Computer program pownor for fitting the power-normal and-lognormal models to life or strength data from specimens of various sizes. *NASA STI/Recon Technical Report N, 92*, 29813.

Neyman, J., & Pearson, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 175-240. http://dx.doi.org/10.2307/2331945

Ramesh, C., & Kirmani, S. (1987). On order relations between reliability measures. *Stochastic Models, 3*(1), 149-156. http://dx.doi.org/10.1080/15326348708807050

Royston, P. (2001). The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Statistica Neerlandica, 55*(1), 89-104. http://dx.doi.org/10.1111/1467-9574.00158

Rutqvist, L. (1985). On the utility of the lognormal model for analysis of breast cancer survival in sweden 1961-1973. *British journal of cancer, 52*(6), 875-883. http://dx.doi.org/10.1038/bjc.1985.272

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6*(2), 461-464. http://dx.doi.org/10.1214/aos/1176344136

Szyszkowicz, S., & Yanikomeroglu, H. (2009). Modified-power-lognormal approximation to the sum of lognormals distribution. *Global Telecommunications Conference, 2009*, 1-6. http://dx.doi.org/10.1109/GLOCOM.2009.5426029

Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 60-62. http://dx.doi.org/10.1214/aoms/1177732360

Modelling Students' Length of Stay at University Using Coxian Phase-Type Distributions

Adele H. Marshall[1], Mariangela Zenga[2] & Sabrina Giordano[3]

[1] Centre for Statistical Science and Operational Research (CenSSOR), Queen's University Belfast, Belfast, Northern Ireland, UK

[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, Italy

[3] Department of Economics, Statistics and Finance, University of Calabria, Arcavacata di Rende, Cosenza, Italy

Correspondence: Mariangela Zenga, Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via B. degli Arcimboldi, Milano 20126, Italy. Tel: 39-2-6448-3158. E-mail: mariangela.zenga@unimib.it

Received: December 17, 2012 Accepted: January 11, 2013 Online Published: January 22, 2013

doi:10.5539/ijsp.v2n1p73 URL: http://dx.doi.org/10.5539/ijsp.v2n1p73

Abstract

The time that Italian students spend at university is remarkably longer than in other European universities. For this reason, the government has recently introduced new rules for academic courses, in order to reduce the issue of long term students. In addition to this, universities need to address the growing problem of students prematurely leaving university before completing their courses. This paper considers the analysis of the length of stay of groups of students at Milano-Bicocca University using Coxian phase-type distributions classified according to the student individual characteristics.

Keywords: Markov chain, classification and regression tree, hazard function, survival function, students progression

1. Introduction and Motivations

Economists and sociologists regard university education as an investment where the costs of the education is balanced against the future benefits of having a better educated population and employable workforce. When a student leaves the university without completing their degree, it is at a cost to the student's family, the university as an institution and society. The costs to society are due to the economic output loss: the graduated is more productive than the non-graduated and society does not make a profit from the taxes of the missed graduate. Following this line of thought, the reform of Italian universities, through Law 509/1999, aims to prevent university drop outs and shorten the time taken to obtain a degree.

Moreover, graduation and drop out rates are adopted as criteria for evaluating the performance of universities. Therefore, the challenge for university managers is to make better informed policy decisions that can streamline the degree completion process, reduce the length of time it takes students to complete a degree and develop effective programmes to prevent drop outs.

The focus of this paper is on the time-to-event data where the time is the number of days elapsed from when a student first enrols until he/she experiences the event, that is, he/she graduates or drops out from that university course.

The analysis of university education has been developing systematically for a long time and a wide literature exists on this subject. Event-History analysis (DesJardin et al., 1999, 2006; Ishitani, 2003; Kalamatianou & McClean, 2003) focuses on discrete events (the student drop out or graduation) occurring over time to establish risk factors. This technique is particularly interesting for analysing the departure process because the assessment of the transition from one state to another, that is, from enrolled to not enrolled, and the identification of the factors (e.g. personal, academic, socio-economic status of family) which influence the students' decision of leaving, are attainable (Triventi et al., 2009).

Another piece of work makes use of Markov chain models to analyze the progression of students at university. In these models, every student occupies a state at time t and transits from state to state at time $t+1$ (the first and the last state represent enrolment and graduation/drop out respectively, while other states represent educational progress).

Gani (1963) originally used a Markov chain for estimating the probability of Australian students completing their degree course. Shah and Burke (1999) provided estimates for the mean time a student takes to complete the course, and mean time students spend in the higher education system in Australia. Harden and Tcheng (1971) built the transition matrices from available historical data. Other examples are given in Song and Chissom (1994), Sah and Degtriarev (2005). Recently, Symeonaki and Kalamatianou (2011) proposed the theory of non-homogeneous Markov systems with fuzzy states for describing student educational progress in Greek universities.

The aim of this paper is to analyse student progression in the Italian reformed degree system and to estimate the influence of various factors on the probability that students, with certain characteristics, will progress successfully towards their degree or drop out. We propose to use the Coxian phase-type distribution for modelling the length of stay (in days) of the students enrolled at University of Milano-Bicocca. Student status has been observed for six academic years, during which time the student can graduate (as of the third year), drop out, change course or university, or can still be enrolled at the end of the observation (considered right-censored in the analysis).

There has been a wealth of literature devoted to investigating the determinants of the propensity of students to drop out of the academic career or to complete their degree programme. Personal characteristics such as gender and age; or individual abilities; income, education and socio-economic status of the family, academic-specific factors (services, quality in teaching, etc), and time-varying variables such as number of passed exams or credits are found to affect student choices at university (Arulampalam et al., 2004; Arulampalam et al., 2005; Checchi & Flabbi, 2006; Johnes, 1990; Light & Strayer, 2000; Robst et al., 1998; Smith & Naylor, 2001; among others).

A recent technical report by a consortium of Italian universities (Almalaurea, 2012) showed that one out of two students, enrolled at university, makes a wrong decision about their own education thus resulting in a high rate of drop out and a low level of satisfaction for graduated students. Motivated by this reason, we consider characteristics of students, known on enrolment, such as individual information (age, gender) and pre-college qualification (high school, mark). In this paper, we investigate the potential of using the Coxian phase-type distributions to give insights to the risk of drop out and graduation, and on the probability to "survive" at university for groups of students sharing common characteristics. It is hoped that by making students aware that, on the basis of their education to date, their background and/or their personal characteristics, they are more likely to have a particular outcome; either to drop out, complete their degree in time, or take up to six years or more to graduate, and that this will help steer students towards an appropriate course choice which meets their expectations.

A classification tree is introduced to identify the different student profiles and, for each profile, we model student's time at university (length of stay, LoS) using different Coxian phase-type distributions. A new student upon identifying to which of the considered groups they belong, can gain valuable perspective on his/her probability of finishing the studies or dropping out, and on how long it should take him/her to complete or give up. Within the fitted Coxian phase-type distributions, each phase could represent a specific stage in academic career or behaviour. These issues are also of interest in assessing efficiency at the system and institutional level.

This enhances the use of the Coxian models for offering university leaders possible insights into the actual needs of change in management and lead universities to develop effective retention programmes and initiatives aimed at reducing drop outs and the time taken to complete the degree.

This paper is organized as follows: Section 2 introduces the Coxian phase-type distribution, Section 3 and 4 report the analysis of length of stay at Milano-Bicocca University using Coxian phase-type distributions for groups of students classified according to their individual characteristics. Section 5 concludes the work and reports on possible future developments.

2. Coxian Phase-Type Distribution

Coxian phase-type distributions (Neuts, 1989) are used to describe the time to absorption of a finite Markov chain in continuous time, where there is a single absorbing state $(n + 1)$ and n ordered transient states or phases. The process starts in the first phase, then moves through sequential phases with the choice of entering the absorbing state at any time. For example, the student career at university can be thought of as a series of transitions through latent phases until an event of leaving university occurs due to graduation, drop out or transfer. Absorption from the first phases would represent the drop out of academic programmes, while absorption from the latest phases would indicate the conclusion for those students who complete a degree.

Let $\{X(t); t \geq 0\}$ be a (latent) Markov chain in continuous time with states $\{1, 2, \dots, n, n + 1\}$, where $\{1, 2, \dots, n\}$ are latent (transient) states of the process and state $(n + 1)$ is the (absorbing) state, and $X(0) = 1$.

For $i = 1, 2, \ldots, n-1$ the probability that a unit moves from one phase to the next one in the time interval δt is

$$\text{Prob}\{X(t+\delta t) = i+1 | X(t) = i\} = \lambda_i \delta t + o(\delta t) \tag{1}$$

and for $i = 1, 2, \ldots, n$ the probability that a unit leaves the system by entering the absorbing state is

$$\text{Prob}\{X(t+\delta t) = n+1 | X(t) = i\} = \mu_i \delta t + o(\delta t) \tag{2}$$

The parameters of the Coxian phase-type distribution, λ_i and μ_i, describe the transition rates through the ordered transient states (from state i to state $i+1$) and the transition rates from the transient states to the absorbing state (from state i to the absorbing state $n+1$), respectively, see Figure 1.

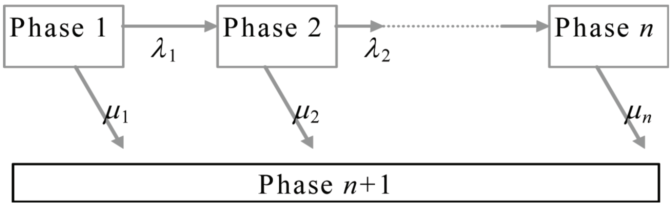


Figure 1. An illustration of the Coxian phase-type distribution

The density and survival functions of the variable T, the time until absorption, are given by:

$$f(t) = \mathbf{p}\exp\{\mathbf{Q}t\}\mathbf{q} \text{ and } S(t) = \mathbf{p}\exp\{\mathbf{Q}t\}\mathbf{1}$$

and the hazard function is $h(t) = f(t)/S(t)$, where $p = (1, 0, 0, ..., 0)$ is the $1 \times n$ vector of probabilities defining the initial transient phases, $q = -Q1 = (\mu_1, \mu_2, ..., \mu_n)'$ is the $n \times 1$ vector of transition rates from transient phases to the absorbing phase, and Q is the matrix of transition rates restricted to the transient phases

$$Q = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & ... & 0 & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & ... & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & ... & -(\lambda_{n-1}+\mu_{n-1}) & \lambda_{n-1} \\ 0 & 0 & 0 & ... & 0 & -\mu_n \end{pmatrix} \tag{3}$$

The probability that the individual leaves the system at phase i, say π_i is determined as a function of the estimated parameters μ_i and λ_i for $i = 1, \ldots, n$, as follows:

$$\begin{aligned} \pi_1 &= \int_0^\infty \mu_1 e^{-(\lambda_1+\mu_1)t} dt = \frac{\mu_1}{\lambda_1+\mu_1}; \\ \pi_2 &= \int_0^\infty \mu_2 e^{-(\lambda_2+\mu_2)t} dt \int_0^\infty \lambda_1 e^{-(\lambda_1+\mu_1)t} dt = \left(\frac{\lambda_1}{\lambda_1+\mu_1}\right)\left(\frac{\mu_2}{\lambda_2+\mu_2}\right); \\ &\vdots \\ \pi_i &= \prod_{l=1}^{i-1}\left(\frac{\lambda_l}{\lambda_l+\mu_l}\right)\left(\frac{\mu_i}{\lambda_i+\mu_i}\right); \\ &\vdots \\ \pi_n &= 1 - \sum_{l=1}^{n-1} \pi_l. \end{aligned}$$

It is a usual procedure to aggregate phases sharing common characteristics to form stages, the interpretation of the stages is, often, more intuitive and meaningful.

Time (length of stay) may then be divided into intervals. In general, the k^{th} length of stay interval (at the k^{th} stage for example) can be determined by $S_k = \left\{t^{(j)} : m\sum_{i=1}^{k-1}\pi_i < j < m\sum_{i=1}^{k}\pi_i\right\}$ where $t^{(1)}, t^{(2)}, ..., t^{(m)}$ represent the ordered lengths of stay data for each individual and m represents the number of observations (Marshall & McClean, 2003).

Parameters μ_i and λ_i, $i = 1, \ldots, n$, are estimated by fitting Coxian phase-type distributions via the EM algorithm (Asmussen et al., 1996) appropriately modified to take censored data into account. Likelihood ratio tests are performed to determine the most suitable number of phases. The likelihood function with censored observations is:

$$L = \prod_{j=1}^{m} f(t_j)^{\alpha_j} S(t_j)^{1-\alpha_j}$$

where α_j is an indicator variable which equals 1, if t_j is a complete time for the j^{th} unit and $\alpha_j = 0$, if t_j is a censored for the j^{th} unit (that is, the event does not occur before the end of the observational period).

Previous research has successfully used Coxian phase-type distributions to represent survival times as the length of time until a certain event occurs, where the phases are considered to be stages in the survival and the absorbing, final stage, the event that occurs causing the individual or element to leave the system completely. For instance, this event could be a patient recovering from an illness, a patient having a relapse, an individual leaving a certain type of employment, a piece of equipment failing, or a patient dying. Faddy (1994) illustrates how useful the Coxian phase-type distributions are in representing survival times for various applications such as the length of treatment spell of control patients in a suicide study, the time prisoners spend on remand and the lifetime of rats used as controls in a study of ageing.

In particular, Faddy and McClean (1999) used the Coxian phase-type distribution to find a suitable distribution for modelling the duration of stay of a group of male geriatric patients in hospital. They found that the phase-type distributions were ideal for measuring the lengths of stay of patients in hospital and showed how it was also possible to consider other variables that may influence the duration. More recently, Marshall and McClean (2003) have demonstrated how the Coxian phase-type distribution can, unlike alternative approaches, adequately model the survival of various groups of elderly patients in hospital uniquely capturing the typical skewed nature of such survival data in the form of a conditional phase-type model (C-Ph) which incorporates a Bayesian network of inter-related variables.

3. Length of Stay at Milano-Bicocca University

3.1 Data

The empirical analysis is conducted on administrative data of students at the Milano-Bicocca University (hereafter MIB). This university was established on June 10, 1998, to serve students from the northern Italy, to relieve some of the pressure on the over-crowded university of Milan, but first of all to offer the opportunity to take a degree at a much more affordable public university than the two renowned private universities of Milan.

The analyzed data refers to just over twenty thousand (20,069) students enrolled in the academic years 2000/01, 2001/02, 2002/3, 2003/04 at MIB in one of the 8 three year degree programmes: Economics (Ec), Educational Science (ED), Law, Mathematics-Physics-Natural Sciences (MPN), Medicine (Med), Psychology (Psy), Sociology (Soc) and Statistics (Stat). Conditions required for admission to the programmes at Milano-Bicocca university vary according to the Faculty. Students who apply for Psychology, Educational Science, Sociology, or to some of the Mathematics-Physics-Natural Sciences degree programmes are selected through an entry test while students of Economics only need to pass a mathematical and Italian language test. In addition, the undergraduate technical courses in Medicine such as Biomedical Laboratory Techniques, Dental Hygiene, Midwifery, Nursing and so on, require students to pass an aptitude test to get enrolled.

For each surveyed student, the time-to-exit from university is measured as the number of days elapsed from when a student first enrols until she/he graduates or drops out or changes institution. Drop out students never finish their degree and are academically dismissed if they provide formal renunciation at any time during the year, if they do not pay taxes or if they do not take exams for at least one year. Students who are still studying but transfer to another institution exit in the data analysis.We follow the performances of the students for six academic years; if a student is still enrolled at the end of the observation time, his/her length of stay at university is considered as right censored in the analysis.

The event of interest is whether the student leaves during their study, gets a degree, transfers to another university, or takes six or more years to finish. The life table of all the students during the six years under study is represented graphically in Figure 2. In the first 3 years, 6167 students (31%) drop out and leave their academic career, 832

(4%) change university and 4675 (23%) take the degree in the regular time. Surprisingly, out of those completing within three years, 34% did so at the beginning of the third year. It is interesting to observe that 4817 students complete the degree programme after the legal duration of the courses. They are known as *fuoricorso* students representing 50.7% of graduated students, which is consistent with the other Italian universities (Miur, 2011). At the end of the observation, 1256 students are still enrolled (6%) taking at least six years to complete their study.

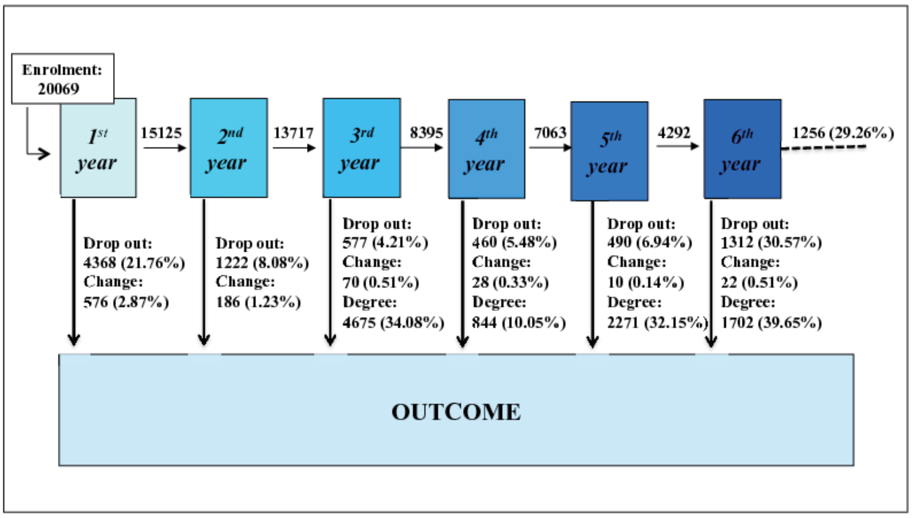


Figure 2. The progression of students during the six years of observations. The percentage is calculated with respect to students enrolled at the beginning of each academic year

As well as the length of stay, various attributes of each student are collected at the time of enrolment: gender, residence (Milan, out of Milan), high school (liceo, technical-vocational-training schools), high school mark (between 0.6 and 1), age at enrolment (at most 19, over 19 years old), enrolment (immediately after the high school, one or more years after the high school), cohort for enrolment (2000/01, 2001/02, 2002/03, 2003/04).

Table 1. Distribution of students by faculties and information at enrolment

		Faculty							
		Econ.	Law	Med.	Psych.	Educ. Science	Math. Physics, Nat. sc.	Stat.	Soc.
Gender	*Female*	46.17%	59.62%	78.28%	74.22%	82.88%	36.55%	50.71%	70.89%
	Male	53.83%	40.38%	21.72%	25.78%	17.12%	63.45%	49.29%	29.11%
Residence	*Milan*	66.95%	66.17%	31.54%	53.19%	54.46%	59.89%	58.40%	52.56%
	Out/Milan	33.05%	33.83%	68.46%	46.81%	45.54%	40.11%	41.60%	47.44%
High Sch.	*Liceo*	34.16%	37.14%	31.30%	57.98%	33.29%	54.04%	51.85%	41.44%
	Other	65.84%	62.86%	68.70%	42.02%	66.71%	45.96%	48.15%	58.56%
Mark	*Average*	0.75	0.74	0.74	0.81	0.75	0.77	0.79	0.79
	Stand. Dev.	0.12	0.11	0.11	0.12	0.11	0.13	0.12	0.12
Age	*<= 19*	57.26%	55.23%	49.88%	59.84%	49.49%	68.73%	70.09%	49.15%
	> 19	42.74%	44.77%	50.12%	40.16%	50.51%	31.27%	29.91%	50.85%
Enrolment	*Immediat.*	75.25%	74.88%	61.46%	70.38%	63.59%	85.14%	81.48%	63.62%
	Later	24.75%	25.12%	38.54%	29.62%	36.41%	14.86%	18.52%	36.38%
Cohort	*2000/01*	24.63%	20.17%	24.78%	25.03%	26.51%	26.89%	20.57%	26.43%
	2001/02	21.82%	24.26%	25.31%	23.67%	25.09%	22.72%	21.51%	25.23%
	2002/03	23.94%	26.05%	24.90%	25.50%	23.82%	23.11%	25.06%	22.21%
	2003/04	29.60%	29.53%	25.01%	25.80%	24.58%	27.28%	32.86%	26.14%
Enrolled Students		***4953***	***1481***	***1243***	***2211***	***3241***	***4772***	***351***	***1817***

Table 1 shows the distribution of students by personal information and Faculty. The composition of the surveyed students is very heterogeneous among the Faculties. In particular, the courses of Mathematics and Physics attract much more male students, while Medicine, Psychology, Sociology and Educational Science have typically a female setting. Most students of Statistics and Mathematics-Physics-Natural Sciences enrol immediately after the high school graduation. In general, more than half students live in Milan, but two out of three students enrolled on Medicine programmes come from other cities, probably because the offered courses are not commonly supplied in other Italian universities.

The preferences of students coming from different types of high school are in one case unexpected. Two-thirds of the students enrolled in Economics, Educational Sciences, Law and Medicine attended professional high schools while half the group of students who prefer to apply for the other programmes are qualified at liceo. It is unexpected that 63% and 68% of enrolled students into Law and Medicine courses respectively are graduated at a technical high school. More or less the same percentage of students enrolled in the four cohorts.

The final mark at the high school (Note: As the type of scale of mark at the high school has changed during the last 10 years, we considered a homogeneous version with minimum value 0.6 and maximum 1) seems not to be a discriminating factor. The average mark is quite similar for all the Faculties, except for the Psychology and Statistics programmes which are preferred by students who graduate with a medium-high level. The smallest average high school mark is for students on Medicine courses.

Table 2 shows the percentage of students of every Faculty by the cause of exit from MIB. Overall, almost half of students (47%) succeed to earn a degree, but a high percentage of enrolled students (42%) do not complete the degree programmes. At the end of the observation time (6 years), 6% of the students are still enrolled. As regards the Faculties, students attending courses of Law and Economics perform the worst. The rate of students who start studying Law (Economics) and never finish is dramatically high, 60% (53%). The Faculty of Law has also the highest percentage (10%) of students still "in progress" toward the degree after six years.

On the other hand, the percentage of students who graduated in Medicine programmes is substantially higher: nearly eighty percent of enrolled students complete their degree, while only 17% of students decide to drop out and less than 1% take over five years to finish. Note that these students have the smallest average mark at high school, but they do better in attaining a degree than their colleagues.

Students enrolled on Educational Sciences, Psychology, Sociology and Statistics courses perform similarly, with approximately 55% of these students completing their degree and nearly 35% dropping out.

Table 2. Distribution of students by Faculties and exit from MIB

		Faculty							
		Econ.	Law	Med.	Psych.	Educ. Science	Math. Physics, Nat. sc.	Stat.	Soc.
Exit	Drop out	52.80%	59.89%	16.49%	35.19%	35.24%	41.60%	35.33%	38.14%
	Degree	34.79%	26.00%	79.40%	55.99%	55.57%	45.66%	55.27%	54.21%
	Transfer	5.69%	3.78%	3.38%	3.98%	1.97%	6.29%	3.42%	2.64%
	Still enroled	6.72%	10.33%	0.72%	4.84%	7.22%	6.45%	5.98%	5.01%
Enroled Students		***4953***	***1481***	***1243***	***2211***	***3241***	***4772***	***351***	***1817***

3.2 A Classification Tree

A CART (Classification and Regression Tree) is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. Specifically, it is a non-parametric tree-structured recursive partitioning method, introduced by Breiman et al. (1984), to predict a response variable on the basis of certain predictors observed on a learning sample. The algorithm consists of two main stages: growing and pruning. In growing, the tree is recursively partitioned into subsets (nodes); each partition is obtained by examining all the possible binary splits along the observed data of each predictor variable and selecting the split that most reduces some measure of node impurity. The result is a sequence of nested trees, with increasing numbers of leaves (terminal nodes), until no more splits are possible and the fully grown tree is reached. The pruning operation on the fully grown tree aims then to select the best subtree and consists by declaring an internal node as terminal and deleting all its descendants; this makes the tree more general and prevents any over-fitting on the training set. The aim of the classification tree is to predict the level of the response on the basis of the vector of the explanatory variables.In this paper the results for CART are obtained using the R package *rpart* (Therneau, 2012).

We created the classification tree reported in Figure 3 to classify the students according to their propensity for completing their degree based on individual characteristics collected at enrolment. Actually, the categorical response variable of interest is the event that determines the exit from MIB with four categories: degree, drop out, change, still enrolled. In the nodes of the tree, students with common attitude towards study who behave in a similar way are joined.

The Gini index is used to evaluate the node impurity and the misclassification rate at the final stage is 0.387. The mode of the four categories is shown on the final nodes in Figure 3. Faculty (1), age at enrolment (1), mark at high school (0.7), enrolment time (0.6) and high school type (0.2) give the most relevant contribution in the classification. The normalized measure of the importance of each predictor variable in relation to the final tree is reported in brackets.

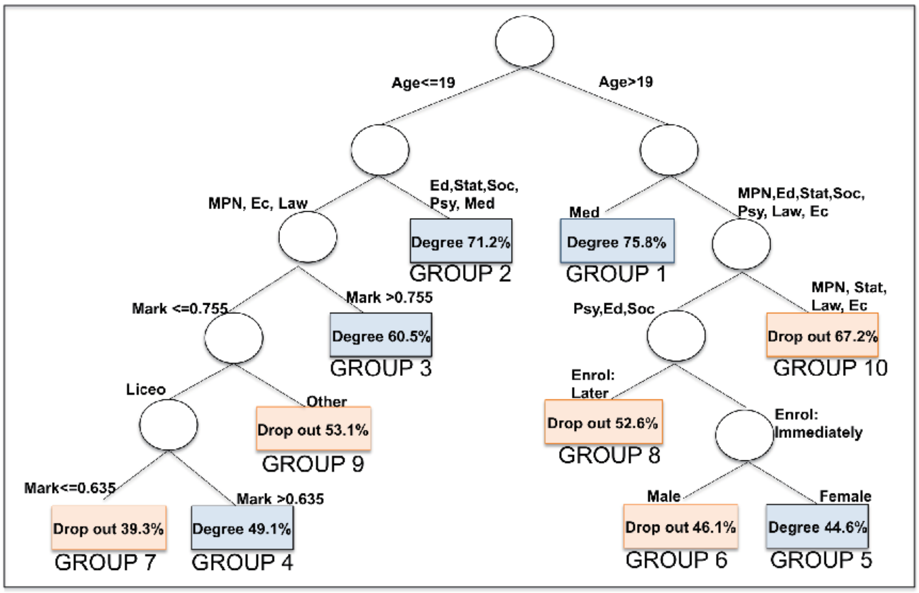


Figure 3. The classification tree for the MIB students by the risk of graduation or drop out

At the end of the pruning procedure, the terminal nodes of the tree identified 10 groups. Table 3 reports the description of the 10 groups ranked according to the likelihood of graduation, the graduation risk. For every group identified, the most frequent cause of exit is reported (column 6, Table 3). The last three columns of the table contain the mean, median and coefficient of variation of the length of stay(in days) at MIB of all the students belonging to each group.

The first group, for example, associated to a high risk of graduation, consists of over 19 years old students in Medicine. These students are probably the most motivated in studying: as they have been forced to pass an admission test and are mostly living outside of Milan, deciding to come to Milan probably only to attend courses. The study programmes give professional training where students interested in such courses make job-oriented choices without wasting time. The students tend to be older than their colleagues probably as most of them have failed the test previously. Their length of stay at MIB on average is 1031 days, so they are likely to complete their degree in a timely manner. At the end of the six years only 2 students are censored.

The last node (hereafter, the tenth), corresponding to the highest risk of drop out (67%) combined with the smallest percentage of graduated students (14%), and refers to the group of students who enrol in Economics, Mathematics-Physics-Natural Sciences, Statistics and Law at least one year later than graduation at school when they are older than 19 and drop out of the academic programmes during the first 2 years (average LoS is 775 days). They resemble the category of students who enter university without any real motivation, probably under the pressure of their family who believe in the usefulness of the degree to getting a job, but who give up along the way.

The ninth group consists of students qualified at high schools different from liceo with at most a medium mark (<=0.775), enrolled immediately after school into Mathematics-Physics-Natural Sciences, Economics and Law. They could be those students who enter college with a weak background and do not perform at the level required to meet the Faculty standards and decide to leave, in fact half of them drop out.

Another group, the third, associated to a high risk of graduation (60%) consists of students enrolled immediately after school with the highest marks at Faculties Mathematics-Physics-Natural Sciences, Economics and Law. Their colleagues of the same Faculties and same age, but graduated at liceo with low-medium marks, compose the fourth group which rises above the others for the longest stay at MIB, 50% of them do not complete degree programmes within normal time and takes more than 3.64 years. Another node (the sixth), corresponding to a medium risk of drop out (46%) combined with a slightly lower percentage of graduated students (36%), refers to the group of male students over 19 years old, enrolled immediately after school into Educational Science, Sociology and Psychology Faculties. They could be those students who prefer to reconcile work with study, the programmes of these Faculties are in fact suitable also to part-time students.

Table 3. Description of the groups identified by the classification tree

Group	Description	N. students	Chance of Graduation	Drop out Risk	Outcome	Mean LoS	Median LoS	Variation Coefficient LoS
1	Age>19; Med	623	75.80%	19.40%	DEGREE	1031.53	1163.00	0.44
2	Age<=19; Fac: Ed, Stat, Soc, Psy, Med	4686	71.10%	17.40%	DEGREE	1230.65	1231.50	0.45
3	Age<=19 Fac: MPN, Ec, Law Mark >0.755	3457	60.50%	21.40%	DEGREE	1195.67	1189.00	0.51
4	Age<=19 0.635<Mark<=0.755 Fac: MPN, Ec, Law High School: Liceo	1223	49.10%	26.10%	DEGREE	1245.10	1358.00	0.56
5	Age>19 Fac: Ed, Soc ,Psy Enrolment: Immediately Gender: Female	668	44.60%	32.30%	DEGREE	1229.69	1320.00/	0.57
6	Age>19 Fac: Ed, Soc, Psy Enrolment: Immediately Gender: Male	360	36.40%	46.10%	DROP OUT	1068.93	1183.00	0.67
7	Age<=19 Mark<=0.635 Fac: MPN, Ec, Law High School: Liceo	496	29.20%	39.30%	DROP OUT	1177.41	1311.00	0.68
8	Age>19 Fac: Ed, Soc, Psy Enrolment: Later	2421	28.70%	52.60%	DROP OUT	1019.69	980.00	0.76
9	Age<=19 Mark<=0.755 Fac: MPN, Ec, Law High School: Prof/Tech	1758	23.90%	53.10%	DROP OUT	991.84	763.00	0.80
10	Age>19 Fac: MPN, Ec, Law, Stat	4377	14.70%	67.2%	DROP OUT	775.16	362.00	0.96

Table 4 reports the percentage of students in each group who enter university and go on to complete degree programmes within normal time (*in corso*) or take at least 4 or more years to finish or dropping out (*fuoricorso*).

Table 4. Rates of *in corso* and *fuoricorso* students by groups

		Groups										***Total***
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Graduation	*In corso*	86.2%	74.5%	64.8%	42.8%	59.1%	51.1%	23.4%	53.3%	31.4%	36.8%	49.3%
	Fuori corso	13.8%	25.5%	35.2%	57.2%	40.9%	48.9%	76.6%	46.7%	68.6%	63.2%	50.7%
	Total	***472***	***3333***	***2093***	***601***	***298***	***131***	***145***	***696***	***420***	***642***	***8831***

Students belonging to the first group are considerably more likely to graduate within three regular years. Most of students in the second and third groups complete their degree in a timely manner. The percentage of students who graduate *in corso* is higher for group five than for group six even if the graduation rate is in the opposite order (see Table 3). The other groups can be characterized by high drop out rates and low graduation rates, having also a high rate of *fuoricorso* students.

3.3 The Survival and Hazard Functions for the 10 Groups of Students

To complete the explorative study of the length of stay at MIB, the empirical survival and hazard functions for the ten groups of students can be determined.

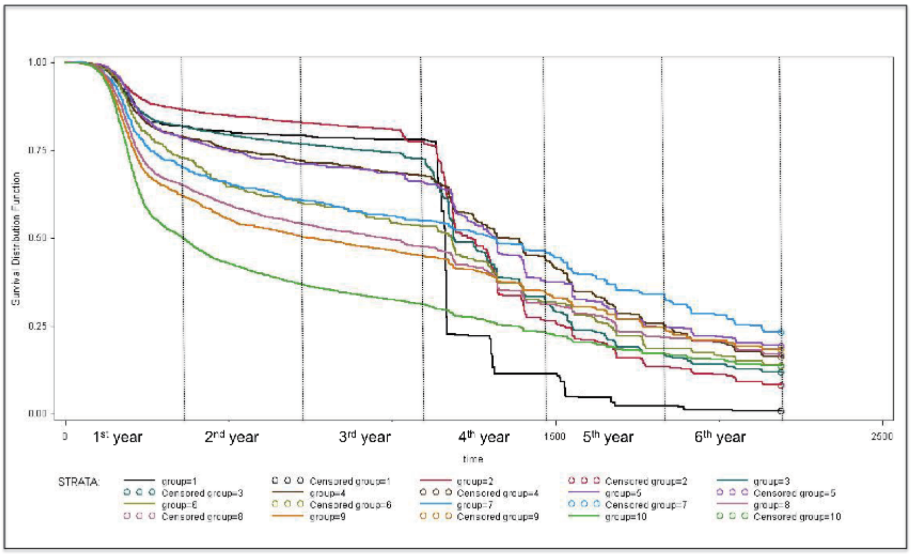


Figure 4. Empirical survival curves for the ten groups of students

Figure 4 displays the empirical survival curves for each group according to the product limit estimator. The shape of the curves seems quite similar: in the first academic year, a steep decline appears overall due to the high rate of drop outs, followed by a stationary trend in the next two years where a reduced number of drop outs and transfers usually registered. Starting from the end of the third year, there is a gradual decrease of the survival probability of students completing their degree.

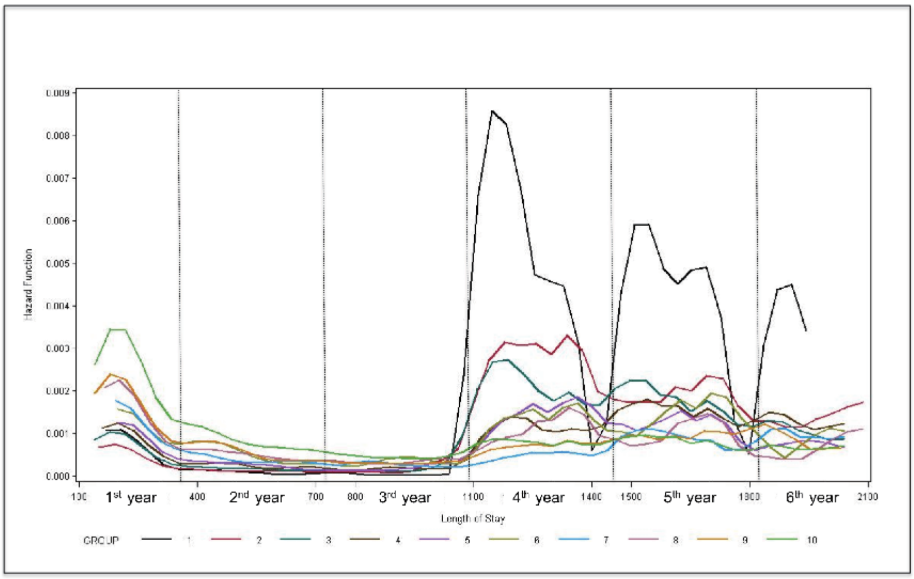


Figure 5. Empirical hazard functions for the ten groups of students

Although the general trend of the survival curves is common to all the groups, the plotted functions do not overlap and the log-rank test confirms that a significant difference exists among them (Chisq=538, df=9, p-value=0.0001).

In particular, the empirical survival curve of the first group stands out from all the other groups. It shows a sudden decline around the end of the third academic year, where most students take the degree. The curve of the second group, instead, dominates the other curves until the graduation time occurs, as the drop out/transfer rate registered for this group of students in the first three academic years is the smallest.

The performance of students decreases as the group number rises.

An overall inspection of Figure 5 shows that the higher the number of group the higher the risk of drop out at the beginning of the academic career, moreover, the first groups are characterized by a greater risk of completing the degree programme then the latest groups. This seems reasonable to expect given that groups one and two have the highest percentage of students completing within three years, therefore the risk of completing should be greater. Likewise, it is in the earlier period of study that you would expect students to be most indecisive of their course choice and most likely to drop out.

4. Fit of the Coxian Phase-Type Distribution

Tables 5a and 5b report the fit of the Coxian phase type distribution parameters using the EM-algorithm (Asmussen et al., 1996) adjusted for censored data. From inspection of the results, it is apparent that a 19 phase Coxian phase-type distribution is the most suitable for all the data (Note: We use a Chi-square test for nested model where the Loglikelihood for 19 phases was -136603.4351 and the Loglikelihood for 20 phases was -136602.9433, p-value=0.3884) together. However, it is important to note that some of the parameters μ_i associated with phase i in Table 5a are equal to zero. This would suggest that no one is observed leaving this phase which can therefore be aggregated with the neighbouring phase. Hereafter, the term stage will indicate a set of sequential phases with estimates of μ_i approaching to zero aggregated together with the closest phase associated to a strictly positive μ_i. In effect, the phases with small values of μ_i parameters are redundant and only the most dominant phases with the largest μ_i values are meaningful. This will also prevent an over-fitted model as reported in earlier literature (Marshall et al., 2012). In each stage we calculated the probability of leaving university due to degree, or drop out, or transfer. Actually, in each of the estimated phases where $\widehat{\mu_i} = 0$ it is possible to leave the process of study, but the probability is so small that it is more likely for students to stay in the process than leave.

Table 5a. Results of fitting Coxian phase-type distribution

Phase i	***Stage***	$\hat{\lambda}_i$	$\hat{\mu}_i$	π_i
1	Explorative	0.0179	0	0.0003
2		0.0177	0	0.0000
3		0.0177	0	0.0000
4		0.0121	0.0056	0.3183
5		0.0143	0	0.0000
6	Intermediate	0.0143	0	0.0000
7		0.0143	0	0.0000
8		0.0143	0	0.0000
9		0.0132	0.0010	0.0488
10	Outcome	0.0116	0	0.0000
11		0.0116	0	0.0000
12		0.0116	0	0.0000
13		0.0116	0	0.0000
14		0.0116	0	0.0000
15		0.0116	0	0.0000
16		0.0116	0	0.0000
17		0.0116	0	0.0000
18		0.0030	0.0086	0.4682
19	Tardive	-	0.0003	0.1644

Table 5b. Average and interval of length of stay at each stage

Stage k	*Number of students (leaving at the stage)*	*Lower bound* LoS	*Upper bound* LoS	*Average of* LoS
1	6394	0	464	217.53
2	980	465	732	584.98
3	9366	733	1916	1326.26
4	3299	1917	2190	2161.23

The length of stay of all the surveyed students at MIB is analysed and found to be most suitably represented by a 4 phase Coxian distribution. The estimated parameters indicate four positive absorption probabilities (Note: Actually the first value of π_i is positive but strongly close to zero and it is ignored). So, the career of students seems to go through four sequential stages which we name explorative, intermediate, outcome and tardive. The four stages appear to represent student behaviour appropriately. At the beginning of a university course (the first explorative stage), we imagine an explorative stage in which the students face a new form of study; some of which realize that they cannot perform at the level required to meet the faculty standards and are discouraged to continue. So, the impact of the new environment results in a peak of students leaving (the drop out students). The second intermediate stage relates to students who have previously been unsure of their career path and after an initial attempt to go on, decide not to pursue their studies further where they either drop out or transfer, other students rest "in progress" towards a degree, proceeding step by step. The third stage, the outcome stage, comprises of the motivated students who complete their degree, in a timely manner however there will also be some students who are 'resting in progress'. The final stage (the fourth, tardive stage) are those students, with the longest length of stay at university (*fuoricorso* or censored data) taking six or more years to complete their degrees.

In particular,

• The explorative stage has length of stay between 17 and 464 days. The mean time to departure from MIB is 218 days, that is, less than three quarters of the first academic year. Among students leaving university at this stage, 88% gives up studying completely, while the remaining 12% decide to transfer to another university.

• The intermediate stage has length of stay interval between 465 and 732 days. Here we see the group of students who have been in doubt with a mean stage of 585 days (about one and half years) on whether keep up the pace of study. Unfortunately, 92% of students who leave university during this stage are drop out students while 8% of them choose to transfer. Only 2 students take the degree in this stage (formally they graduated in the first degree session, on June).

• The outcome stage has length of stay between 733 and 1916 days. This comprises strongly motivated students who complete their degree. The average time it takes students to earn the degree in this phase is 1236 days corresponding to about three academic years. In particular, 87 out of 100 students will exit their academic career at this stage and graduate within the regular time, the other students drop out.

• The tardive stage has length of stay varying between 1917 and 2190 days. This final stage is regarded as the tail period where *fuoricorso* students graduate or remain still enrolled after six years (censored LoS). Eighty percent of students complete their degree while 20% percent remain enroled up to the end of the period. Students leaving at this stage belong to the group of those who wish to graduate but they succeed only after a very long stay at university. The average time to degree is equal to 2161 days, about six years, so twice the regular duration of a degree programme. Understanding the factors that are influential to such a long degree completion time is one of the most crucial issues for the university managers.

Table 5b reports the students leaving the university, the interval and the average LoS for each stage. So, for example, 6394 students leave during the explorative stage within 464 days and their career lasts on average 217.53 days.

In describing the stages, we focused on students who left the system, but of course there are all the motivated students who rest in progress toward a degree and proceed through these sequential stages increasing their abilities (and their human capital) by attending courses and passing exams.

Figure 6 displays how the estimated Coxian phase-type distribution fits the empirical distribution of the lengths-of-stay at MIB university. The fitted density seems to meet the characteristic shape with two peaks, the former due to the high drop out rate and the latter related to students graduating with a degree.

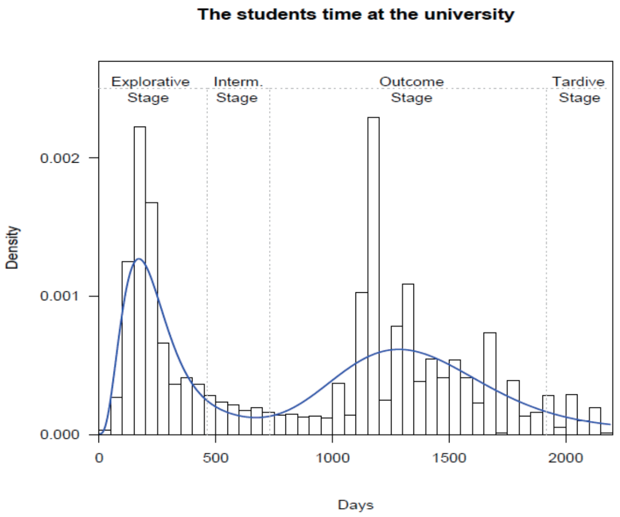


Figure 6. Empirical and estimated distribution of the length of stay at MIB university (on the complete sample of students)

At this point, we investigate the fit of a Coxian phase-type distribution to model the length of stay at MIB for each group of student. Table 7 reports the results of the fitting procedure. The estimated phases (first row) are then aggregated to form stages considering the positive values of μ_i. The likelihood statistic and p-values show that there is no significant improvement in fit by adding one more phase to the distribution. The final rows indicate the bounds (in days) of the intervals of the length of stay of students leaving at every stage of the distribution. So, for example, the LoS of students in the 10th group are divided into 3 intervals (delimiting the 3 stages): 0-412, 413-745, 746-2190 (days).

At first glance, the results make it clear that the number of stages is quite different within the different groups. Only the distribution for the second group has the 4 stages detected in the distribution fitted on the complete sample of students. Moreover, for students in groups 3 to 6, the intermediate stage seems not to be relevant.

A case which deserves attention is the first group, in such a case a remarkable high number of phases are registered and the algorithm for parameter estimation appears to have difficulty in reaching convergence. It is possible that a more suitable mixed model (continuous-discrete) for the student performances of this group can avoid the convergence difficulty. This problem needs further investigation.

A comparison between the empirical and estimated distributions for each group is shown in Figure 7. The Coxian phase-type distributions differ according to the different student groups but almost all the fitted densities seem to suitably represent the empirical trends, even if for some groups the fitting is substantially better than for other groups.

Table 7. Results of fitted Coxian distribution to the length of stay in the 10 groups of students

	Groups									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
# Phases	32	22	14	12	11	11	13	12	10	12
Log likelihood	-4338.0	-32799.4	-23787.2	-8225.2	-4356.1	-2461.9	-3085.9	-15865.1	-11295.5	-28277.6
Test Statistics	4.57	1.22	1.21	3.82	0.84	0.92	2.45	4.88	3.13	3.90
p-value	0.05	0.27	0.27	0.07	0.33	0.31	0.15	0.04	0.10	0.07
#$\mu_i > 0$ (n. of stages)	3*	4	2*	2	2	2	3	3	3*	3*
up. bound 1	382	464	585	663	568	606	403	460	417	412
up. bound 2	732	653	2190	2190	2190	2190	866	706	732	745
up. bound 3	2190	1889	-	-	-	-	2190	2190	2190	2190
up. bound 4	-	2190	-	-	-	-	-	-	-	-

Description: the* indicates that actually one more value of μ_i is positive but strongly close to zero and it is ignored. The upper bound (up. bound in the table) of the intervals of LoS at each stage is indicated, the first lower bound is zero.

Recall that the groups are ordered according to the chance of graduation, so the later groups involve students whose performance is poor and are more likely to drop out than their colleagues of the first groups with a marked propensity towards study. In Figure 7, the higher the number of group, the better the fit of the Coxian phase-type model, thus the use of the Coxian phase-type distribution seems to be more appropriate for modeling the MIB of students who perform worse and have very long lengths of stay. This agrees with previous research where Coxian phase-type distributions are used to represent survival or length of stay of elderly patients in hospital. There is more heterogeneity in the earlier stages of survival which is to be expected as there is a bigger case mix of individuals present in the first phases. In fact Faddy and McClean (1999) and Marshall et al. (2003) both highlight that the first phase includes elderly patients who either leave the system quite quickly due to having minor problems and thus return home or who have critical health problems and die within a short period of time in hospital. The approach is good at representing the very long stay patients who are consuming large amounts of hospital resources by staying in hospital for a long period of time. Likewise, the research presented in this paper is primarily concerned with those students who have very long stays at University and do not complete their course within six years of stay.

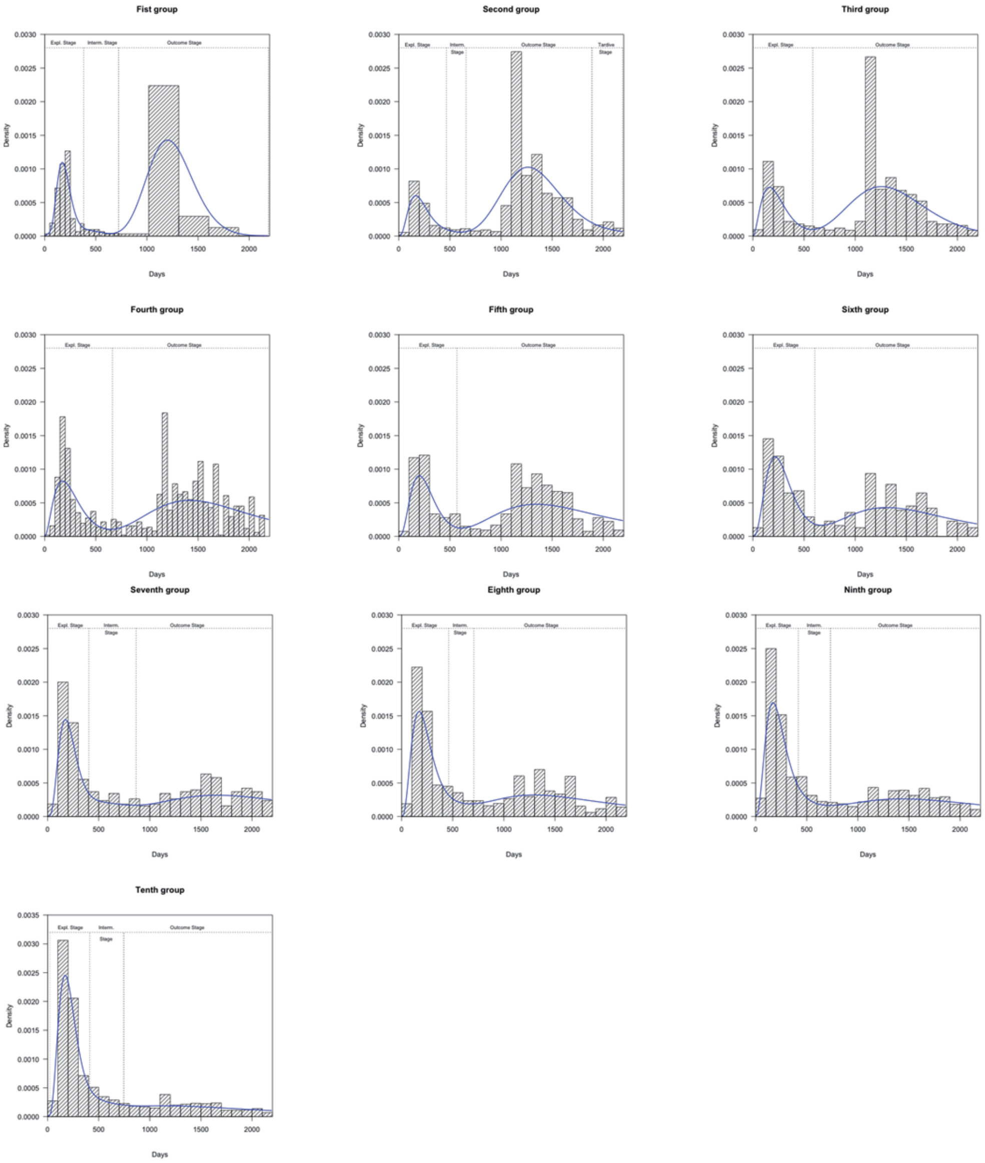


Figure 7. The empirical and estimated density distributions for the ten groups

As expected, at first sight, comparing the Coxian phase-type distributions in succession from the first to the latest ones, it appears clear that the second peak in the distributions, relates to lengths of stay of graduated students, gradually tailing off as the number of group increases, while the peak in the initial stage caused by the dropping out students rises. Thus, the distributions are initially (for the first groups) bimodal and then tend to become highly skewed with only one large peak at the explorative stage (for the last groups).

In the first 4 plots, students are likely to enter the absorbing state either for dropping out and for graduation. Plots relative to all the groups from the 5th to the 10th instead show that most of the students reach the absorbing state in the first explorative stage, a high rate of enrolled students never finish their degree. This represents the very challenge of university leaders who need to ensure policies and practices to prevent this academic failure.

The percentage of students who graduated is substantially higher for the first three groups of students, the corresponding plots exhibit the largest peak of the density at the outcome stage. In the third group, for example, students with a good background at school who decide to enroll straight into the academic programmes of Mathematics-Physics-Natural Sciences, Economics and Law are more likely to complete their degree. The Coxian phase-type distribution captures this performance. On the other hand, the over 19 year old students in the eight group, enroled in Education, Sociology and Psychology Faculties at least one year after the high school, do not overcome the initial difficulties and most give up in the first two years (approximately 460 days). The Coxian phase-type distribution is able to represent this empirical mode in the explorative stage (see Figure 9).

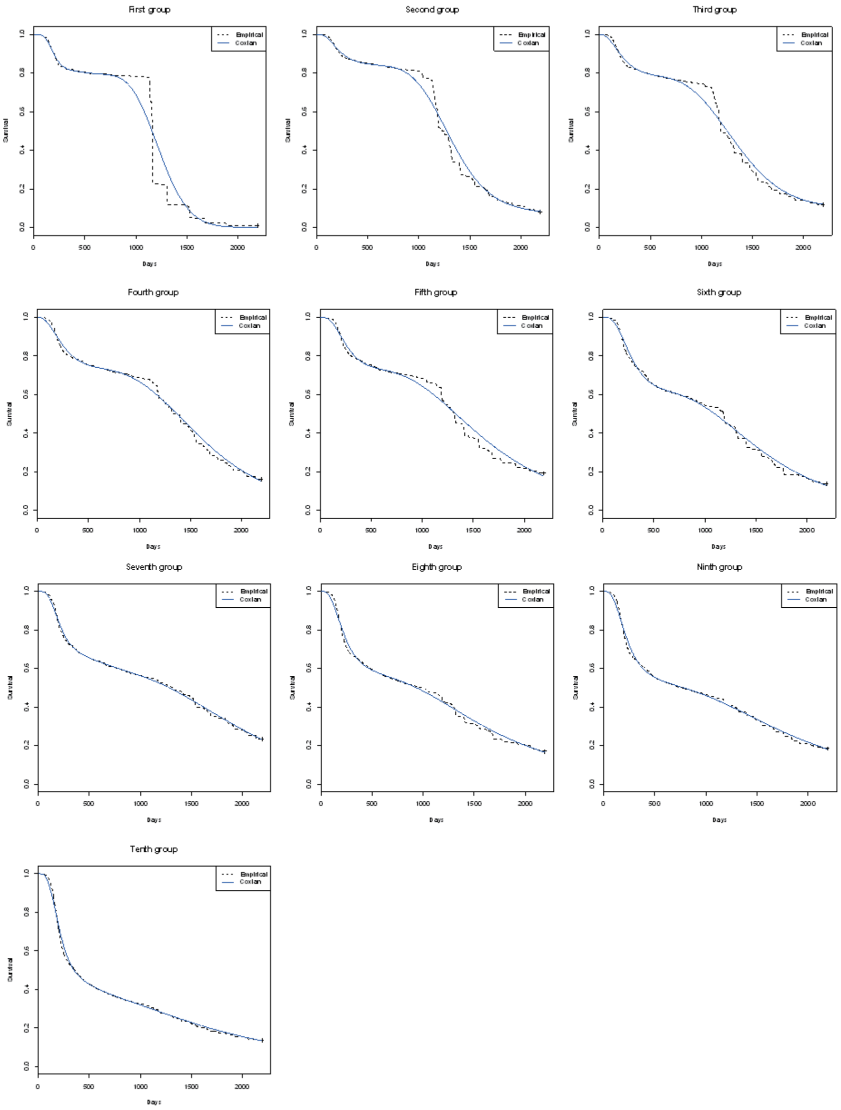


Figure 8. The empirical and estimated survivals for the ten groups

For the first group the fitted distribution does not capture the extremely high second peak of the empirical distribution due to the fixed graduation dates where graduation falls on fixed days in the academic year. An alternative representation is to consider a mixture of Coxian phase-type distributions. However, doing so does not improve the fit any further than what is presented in this paper. This aspect will undergo further investigation and we restrict the focus of this paper on the extremely long stay students.

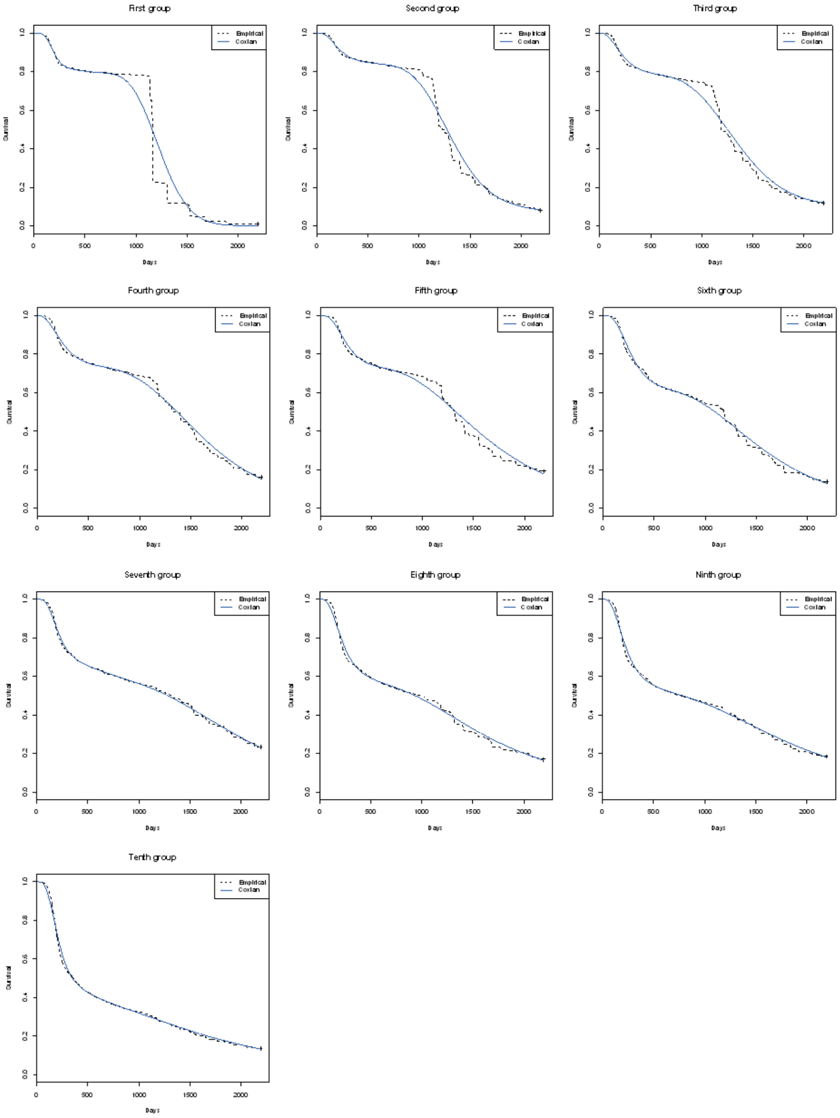


Figure 9. The empirical and estimated hazard for the ten groups

Fitting the Coxian phase-type models enables us to offer possible insights on estimating the risk of leaving university due to drop out or graduation reasons according to the group to which the student belongs. The Coxian phase-type distribution also provides the estimates of the probability to "survive" in the university system towards the degree and distinguish between different groups of students depending on the survival probability they have.

Survival and hazard functions are estimated using the Coxian phase-type densities of each group, and are compared with the empirical curves determined by the non-parametric Kaplan-Meier procedure in Figures 8 and 9, respectively. The estimated functions approximate the empirical curves well, where the fitted results for the densities represent the highest ranked groups most appropriately (Figure 7). In particular, the estimated survival functions overlap the empirical ones for all of the later groups (particularly for groups 7-10).

5. Conclusion

The work presented in this paper introduces an innovative application of the Coxian phase-type distribution to the University student progression and drop out phenomena. There are two models considered. The first, introduces a classification tree to divide the students into different profiles of stay according to their characteristics known on enrolment of their course. This produced ten groups of student with differing characteristics across the groups and time at university represented using different Coxian phase-type distributions. A new student, upon identifying the group to which he/she belongs, can gain valuable perspective on his/her probability of finishing the studies or dropping out, and on how long it should take him/her to complete or give up. Within the fitted Coxian phase-type distributions, each phase represents a specific stage in academic career or behaviour. These issues are also of interest in assessing efficiency at the system and institutional level. This enhances the use of the Coxian models that would offer university leaders possible insights into the actual needs of change in management and lead universities to develop effective retention programmes and initiatives aimed at reducing drop outs and reducing the times taken to complete a degree. Upon developing a model for the ten different student groups, the Coxian phase-type distribution was fitted again separately for each group of student. This provides further refinement of the student length of stay by modeling each student group as a sequence of phases in a separate Coxian phase-type distribution. In doing so, an improvement in survival predictions can be made to the student stay.

This second model follows a similar format to that by Harper et al. (2012) and Marshall et al. (2012) who introduce the Discrete Conditional Phase-type distribution using a classification tree to model patient characteristics on admission to hospital as the first component in the model which is conditioned on the second component, the patient length of stay in hospital represented by a Coxian phase-type distribution. Such an approach is very applicable to student time at University and consistent with previous research. This paper extends that work to another application area and in doing so is able to use the fitted Coxian phase-type distribution to define four stages of student behaviour in University. Linked with these stages different student characteristics and associated likely result for that student in terms of graduating on time or dropping out. Student progression at an University is a concern for many countries particularly the costs incurred and the stress to the student. As further work, it is planned that the models presented in this paper will be applied to student data for other countries. One particular example that will be considered is the application of the model for University students in Greece. Another possible extension to this work is to incorporate the costs into the model.

References

Almalaurea. (2012). XIV Indagine sulla Condizione occupazionale dei laureati. Retrieve from http://www.almalaurea.it/universita/occupazione/occupazione10/

Arulampalam, W., Naylor, R. A., & Smith, J. P. (2004). A hazard model of the probability of medical school drop out in the UK. *Journal of Royal Statistical Society Series A, 167*, 157-178. http://dx.doi.org/10.1046/j.0964-1998.2003.00717.x

Arulampalam, W., Naylor, R. A., & Smith, J. P. (2005). Effects of in-class variation and student rankon the probability of withdrawal: cross-section and time series analysis of UK universities students. *Economics of Education Review, 24*, 251-262. http://dx.doi.org/10.1016/j.econedurev.2004.05.007

Asmussen, S., Nerman, O., & Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics, 23*, 419-441.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Tree*. Chapman & Hall.

Checchi, D., & Flabbi, L. (2006). Intergenerational Mobility and Schooling Decisions in Italy and Germany, *Mimeo*, University of Milan.

Cox, D. R., & Miller, H. D. (1965). *The theory of stochastic processes*. London: Chapman.

DesJardinis, S. L., Ahlburg, D. A., & McCall, B. P. (1999). An event history model of student departure. *Economics of Education Review ,18*(3), 375-390. http://dx.doi.org/10.1016/S0272-7757(98)00049-1

DesJardinis, S. L., Ahlburg, D. A., & McCall, B. P. (2006). The effects of interrupted enrolment on graduation from college: Racial, income and ability differences. *Economics of Education Review, 25*(6), 575-590. http://dx.doi.org/10.1016/j.econedurev.2005.06.002

Faddy, M. (1994). Examples of fitting structured phase-type distribution. *Applied Stochastic models and Data Analysis, 10*, 247-255. http://dx.doi.org/10.1002/asm.3150100403

Faddy, M., & McClean, S. I. (1999). Analysing data on lengths of stay of hospital patients using phase-type distribution. *Applied Stochastic Models in Business and Industry*, 311-317.

Gani, A. (1963). Formulae for projecting Enrollments and Degrees awarded in Universities. *Journal of the Royal Statistical Society, 126*(3), 400-409. http://dx.doi.org/10.2307/2982224

Harper, P. R., Knight, V. A., & Marshall, A. H. (2012). Discrete Conditional Phase-type models utilising classification trees: Application to modelling health service capacities. *European Journal of Operational Research, 291*, 522-530. http://dx.doi.org/10.1016/j.ejor.2011.10.035

Ishitani, T. T. (2003). A longitudinal approach to assessing attrition behavior among first-generation students: Time-Varying Effects of Pre-College Characteristics. *Research in Higher Education, 44*(4), 433-449. http://dx.doi.org/10.1023/A:1024284932709

Johnes, G. (1990). Determinants of student wastage in higher education. *Studies in Higher Education, 15*, 87-99. http://dx.doi.org/0.1080/03075079012331377611

Kalamatianou, A. G., & McClean, S. (2003). The Perpetual Student: Modeling Duration of Undergraduate Studies Based on Lifetime-Type Education Data. *Lifetime Data Analysis, 9*, 311-330. http://dx.doi.org/10.1023/B:LIDA.0000012419.98989.d4

Light, A., & Strayer, W. (2000). Determinants of college completion: school quality or student ability? *Journal of Human Resources, 35*, 299-332. http://dx.doi.org/10.2307/146327

Marshall, A. H., & McClean, S. I. (2003). Conditional phase-type distributions for modelling patient length of stay in hospital. *International Transactions in Operational Research, 10*, 565-576. http://dx.doi.org/10.1111/1475-3995.00428

Marshall, A. H., Payne, K., & Cairns K. J. (2012). Modelling the development of late onset sepsis and length of stay using discrete conditional survival models with a classification tree component. *Proceedings of IEEE Computer Based Medical Systems 2012*. http://dx.doi.org/10.1109/CBMS.2012.6266407

Marshall, A. H., & Zenga, M. (2012). Experimenting with the Coxian Phase-Type distribution to Uncover Suitable Fits. *Methodology and computing in Applied probability, 14*, 71-86. http://dx.doi.org /10.1007/s11009-010-9174-y

Miur. (2011). Undicesimo Rapporto sullo Stato del Sistema Universitario. Retrieved from http://www.cnvsu.it/publidoc/datistat/default.asp?id_documento_padre=11777

Neuts, M. (1989). Structured Stochastic Matrices of M/G/1 Type and Their Application (Marcel Dekker, NY).

Robst, J., Keil, J., & Russo, D. (1998). The effect of gender composition of faculty on student retention. *Economics of Education Review, 17*, 429-438. http://dx.doi.org/10.1016/S0272-7757(97)00049-6

Sah, M., & Degtiarev, K. (2005). Forecasting enrollment model based on First-Order Fuzzy time series, World Academy of Science, Engineering and Technology I , 132-135.

Schultz, T. W. (1963). *The economic value of education*. New York: Columbia University Press.

Shah, C., & Burke, G. (1999). An Undergraduate Student Flow Model: Australian Higher Education. *Higher Education, 37*, 359-375. http://dx.doi.org/10.1023/A:1003765222250

Smith, J. P., & Robin, A. N. (2001). Dropping out of university: a statistical analysis of the probability of withdrawal for UK university students. *Journal of the Royal Statistical Society. Series A, 164*(Part 2), 389-405. http://dx.doi.org/10.1111/1467-985X.00209

Song, Q. & Chissom, B. S. (1994). Forecasting enrollments with fuzzy time series. *Fuzzy Sets Syst., 62*, 1-8. http://dx.doi.org/10.1016/0165-0114(94)90067-1

Symeonaki, M., & Kalamatianou, A. (2011). Markov Systems with Fuzzy States for Describing Students' Educational Progress in Greek Universities, ISI 2011. Retrieved from http://isi2011.congressplanner.eu/showabstract.php?congress=ISI2011&id=1896

Therneau, T., & Atkinson, B. (2012). Package 'rpart'. Retrieve from http://cran.rproject.org/web/packages/rpart/rpart.pdf

Triventi, M., & Trivellato, P. (2009). Partecipation, Performance and Ineqality in Italian Higher Education in 20th Century. *Higher Education, 57*(6), 681-702. http://dx.doi.org /10.1007/s10734-008-9170-0

International Journal of Statistics and Probability; Vol. 2, No. 1; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data

David J. Olive[1]

[1] Department of Mathematics, Southern Illinois University, Carbondale, IL, USA

Correspondence: David J. Olive, Department of Mathematics, Southern Illinois University, Carbondale, IL., USA. Tel: 1-618-453-6566. E-mail: dolive@siu.edu

Received: October 24, 2012 Accepted: January 13, 2013 Online Published: January 22, 2013

doi:10.5539/ijsp.v2n1p90 URL: http://dx.doi.org/10.5539/ijsp.v2n1p90

The research is financed by National Science Foundation grant DMS 0600933

Abstract

This paper presents asymptotically optimal prediction intervals and prediction regions. The prediction intervals are for a future response Y_f given a $p \times 1$ vector $\boldsymbol{x}_f$ of predictors when the regression model has the form $Y_i = m(\boldsymbol{x}_i) + e_i$ where m is a function of $\boldsymbol{x}_i$ and the errors e_i are iid from a continuous unimodal distribution. The prediction intervals have coverage near or higher than the nominal coverage for many techniques even for moderate sample size n, say $n > 10$(model degrees of freedom). The prediction regions are for a future vector of measurements $\boldsymbol{x}_f$ from a multivariate distribution. The nonparametric prediction region developed in this paper has correct asymptotic coverage if the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a distribution with a nonsingular covariance matrix. For many distributions, this prediction region appears to have good coverage for $n > 20p$, and this region is asymptotically optimal on a large class of elliptically contoured distributions. Hence the prediction intervals and regions perform well for moderate sample sizes as well as asymptotically.

Keywords: additive models, nonlinear regression, prediction intervals, prediction regions, regression

1. Introduction

This paper presents asymptotically optimal prediction intervals and prediction regions. The prediction regions are for a future vector of measurements $\boldsymbol{x}_f$ from a multivariate distribution, and are asymptotically optimal on a large class of elliptically contoured distributions. Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $p \times 1$ vector of predictors $\boldsymbol{x}$. The prediction intervals are for a future response Y_f given a vector $\boldsymbol{x}_f$ of predictors when the regression model has the form

$$Y_i = m(\boldsymbol{x}_i) + e_i \tag{1}$$

for $i = 1, ..., n$ where m is a function of $\boldsymbol{x}_i$ and the errors e_i are iid from a continuous unimodal distribution. Many of the most important regression models have this form, including the multiple linear regression model and many time series, nonlinear, nonparametric and semiparametric models. If $\hat{m}$ is an estimator of m, then the ith residual is $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$.

Olive (2007) showed how to form asymptotically optimal prediction intervals for model (1), but for many regression models and estimators, large n is needed for the intervals to perform well. Prediction intervals derived for multiple linear regression did perform well. This paper derives asymptotically optimal prediction intervals that perform well for many models for moderate n.

A large sample $100(1 - \delta)\%$ prediction interval (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \overset{P}{\to} 1 - \delta$ as the sample size $n \to \infty$. Following Olive (2007), let ξ_δ be the δ percentile of the error e, i.e., $P(e \le \xi_\delta) = \delta$. Let $\hat{\xi}_\delta$ be the sample δ percentile of the residuals. Consider predicting a future observation Y_f given a vector of predictors $\boldsymbol{x}_f$ where $(Y_f, \boldsymbol{x}_f)$ comes from the same population as the past data $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Let $1 - \delta_2 - \delta_1 = 1 - \delta$ with $0 < \delta < 1$ and $\delta_1 < 1 - \delta_2$ where $0 < \delta_i < 1$. Then $P[Y_f \in (m(\boldsymbol{x}_f) + \xi_{\delta_1}, m(\boldsymbol{x}_f) + \xi_{1-\delta_2})] = 1 - \delta$.

Assume that $\hat{m}$ is consistent: $\hat{m}(\boldsymbol{x}) \overset{P}{\to} m(\boldsymbol{x})$ as $n \to \infty$. Then $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) \overset{P}{\to} Y_i - m(\boldsymbol{x}_i) = e_i$ and, under “mild” regularity conditions, $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$. If $a_n \xrightarrow{P} 1$ and $b_n \xrightarrow{P} 1$, then

$$(\hat{L}_n, \hat{U}_n) = (\hat{m}(\boldsymbol{x}_f) + a_n \hat{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{1-\delta_2}) \tag{2}$$

is a large sample $100(1-\delta)\%$ PI for Y_f.

According to regression folklore, the percentiles of the residuals are consistent estimators, $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$, under "mild" regularity conditions, and this consistency is the basis for using QQ plots. The folklore is true for linear models: sufficient conditions are $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ and the $\boldsymbol{x}_i$ are bounded in probability. See Olive and Hawkins (2003), Welsh (1986) and Rousseeuw and Leroy (1987, p. 128).

Consider the multiple linear regression model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown iid zero mean errors e_i with variance σ^2. Let the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the ith diagonal element of $\boldsymbol{H}$ for $i = 1, ..., n$. Then h_i is called the ith *leverage* and $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Suppose new data is to be collected with predictor vector $\boldsymbol{x}_f$. Then the leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$.

For the multiple linear regression model, let $\hat{\xi}_\delta$ be the sample quantile of the residuals. Following Olive (2007), let

$$a_n = b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1+h_f)}. \tag{3}$$

Then a large sample semiparametric $100(1-\delta)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n \hat{\xi}_{\delta/2}, \hat{Y}_f + a_n \hat{\xi}_{1-\delta/2}). \tag{4}$$

A PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. The PI (4) is asymptotically optimal on a large class of unimodal continuous symmetric error distributions. For more general distributions, an asymptotically optimal PI can be created by applying the shorth(c) estimator to the residuals where $c = \lceil n(1-\delta) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. See Grübel (1988). That is, let $r_{(1)}, ..., r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$ correspond to the interval with the smallest length. Following Olive (2007), a 100 $(1-\delta)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n \tilde{\xi}_{\delta_1}, \hat{Y}_f + a_n \tilde{\xi}_{1-\delta_2}) \tag{5}$$

where a_n is given by (3). This prediction interval performs well for moderate n for multiple linear regression and several estimators, including least squares.

A problem with prediction intervals is choosing a_n and b_n so that the intervals have short length and coverage close to or higher than the nominal coverage for a wide variety of regression models when n is moderate. Section 2.1 shows how to modify (4) and (5) to achieve these goals while Section 2.2 covers prediction regions for a future vector of measurements $\boldsymbol{x}_f$. Examples and simulations are in Section 3.

2. Method

The idea for finding the asymptotically optimal prediction intervals and regions is simple. Find the target population $100(1-\delta)\%$ covering region. For small n, the coverage of the training data will be higher than that for the future case to be predicted. In simulations for a large group of models and distributions, the undercoverage could be as high as $\min(0.05, \delta/2)$. Let $q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta p/n), \quad \text{otherwise.} \tag{6}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Then use the prediction interval or region that covers $100q_n\%$ of the training data. The coverage of the training data is $100q_n\%$ and converges to $100(1-\delta)\%$ as $n \to \infty$, even if the model assumptions fail to hold.

2.1 Asymptotically Optimal Prediction Intervals

The technique used to produce asymptotically optimal PIs that perform well for moderate samples is simple. Find $\hat{Y}_f$ and the residuals from the regression model. Since the leverage of $\boldsymbol{x}_i$ is closely related to the Mahalanobis distance of $\boldsymbol{x}_i$ from the sample mean $\overline{\boldsymbol{x}}$ of the n predictor vectors, leverage and extrapolation are useful for a wide range of regression models. For a wide range of regression models, extrapolation occurs if $h_f > 2p/n$: if $\boldsymbol{x}_f$ is too far from the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, then the model may not hold and prediction can be arbitrarily bad. This result suggests replacing (3) by

$$a_n = b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2p}{n-p}}. \tag{7}$$

Let $\delta_n = 1 - q_n$ where q_n is given by (6). Then

$$(\hat{L}_n, \hat{U}_n) = (\hat{m}(\boldsymbol{x}_f) + b_n\hat{\xi}_{\delta_n/2}, \hat{m}(\boldsymbol{x}_f) + b_n\hat{\xi}_{1-\delta_n/2}) \tag{8}$$

is a large sample $100(1-\delta)$% PI for Y_f that is similar to (2) and (4).

Let $c = \lceil nq_n \rceil$. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$ correspond to the interval with the smallest length. Then the asymptotically optimal 100 $(1-\delta)$% large sample PI for Y_f is

$$(\hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{1-\delta_2}), \tag{9}$$

and is similar to (5).

To see that the PI (9) is asymptotically optimal, assume that the sample percentiles of the residuals converge to the population percentiles of the iid unimodal errors: $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$. Also assume that the population shorth $(\xi_{\delta_1}, \xi_{1-\delta_2})$ is unique and has length L. Since $b_n \to 1$, $\hat{m}(\boldsymbol{x}_f) \xrightarrow{P} m(\boldsymbol{x}_f)$, and $q_n = 1 - \delta$ for large enough n, it is enough to show that the shorth of the residuals converges to the population shorth of the e_i: $(\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}) \xrightarrow{P} (\xi_{\delta_1}, \xi_{1-\delta_2})$. Let L_n be the length of $(\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$. Let $0 < \tau < 1$ and $0 < \epsilon < L$ be arbitrary. Assume n is large enough so that $q_n = 1 - \delta$. Then $P(L_n > L + \epsilon) \to 0$ since $(\hat{\xi}_{\delta_1}, \hat{\xi}_{1-\delta_2})$ covers 100 $(1-\delta)$% of the data and $L_n = \tilde{\xi}_{1-\delta_2} - \tilde{\xi}_{\delta_1} \leq \hat{\xi}_{1-\delta_2} - \hat{\xi}_{\delta_1} \xrightarrow{P} L$ as $n \to \infty$ since the sample percentiles are consistent and the shorth is the smallest interval covering 100 $(1-\delta)$% of the data. If $P(L_n < L - \epsilon) > \tau$ eventually, then the shorth is an interval covering 100 $(1-\delta)$% of the cases that is shorter than the population shorth with positive probability τ. Hence at least one of $\hat{\xi}_{1-\delta_2}$ or $\hat{\xi}_{\delta_1}$ would not converge, a contradiction. Since ϵ and τ were arbitrary, $L_n \xrightarrow{P} L$. If $P(\tilde{\xi}_{\delta_1} < \xi_{\delta_1} - \epsilon) > \tau$ eventually, then $P(\tilde{\xi}_{1-\delta_2} < \xi_{1-\delta_2} - \epsilon/2) > \tau$ eventually since $L_n = \tilde{\xi}_{1-\delta_2} - \tilde{\xi}_{\delta_1} \xrightarrow{P} L = \xi_{1-\delta_2} - \xi_{\delta_1}$. But such an interval (of length going to L in probability with left endpoint less than $\xi_{\delta_1} - \epsilon$ and right endpoint less than $\xi_{1-\delta_2} - \epsilon/2$) contains more than $100(1-\delta)$% of the cases with probability going to one since the population shorth is the unique shortest interval covering $100(1-\delta)$% of the mass. Hence there is an interval covering $100(1-\delta)$% of the cases that is shorter than the shorth, with probability going to one, a contradiction. The case $P(\tilde{\xi}_{\delta_1} > \xi_{\delta_1} + \epsilon) > \tau$ can be handled similarly. Since ϵ and τ were arbitrary, $\tilde{\xi}_{\delta_1} \xrightarrow{P} \xi_{\delta_1}$. The proof that $\tilde{\xi}_{1-\delta_2} \xrightarrow{P} \xi_{1-\delta_2}$ is similar.

The above results show that PI (9) and the shorth of the residuals behave well when the sample percentiles are consistent. Even if these assumptions do not hold, the PI covers $100q_n$% of the training data, and often the coverage of the future case will be close to $100(1-\delta)$ if the future case Y_f is similar to the training data.

For asymptotic optimality, can not have extrapolation. Also, even if the coverage converges to the nominal coverage, the length of the PI need not be asymptotically shortest unless the highest $1-\delta$ density region of the probability density function of the iid errors is an interval. The highest density region is an interval for unimodal distributions, but need not be an interval for multimodal distributions for all δ. Also see Cai, Tian, Solomon and Wei (2008).

Notice that the technique computes a PI for coverage $q_n \geq 1 - \delta$ which converges to the nominal coverage $1 - \delta$ as $n \to \infty$. Suppose $n \leq 20p$. Then the nominal 95% PI uses $q_n = 0.975$ while the nominal 50% PI uses $q_n = 0.55$. Prediction distributions depend both on the error distribution and on the variability of the estimator $\hat{m}$. This variability is typically unknown but converges to 0 as $n \to \infty$. Also, residuals tend to underestimate the errors for small n. For small n, ignoring estimator variability and using $q_n = 1 - \delta$ resulted in undercoverage as high as $\min(0.05, \delta/2)$. Letting the "coverage" q_n decrease to the nominal coverage $1 - \delta$ inflates the length of the PI for small n, compensating for the unknown variability of $\hat{m}$.

The geometry of the "asymptotically optimal prediction region" is simple. The region is the area between two parallel lines with unit slope. Consider a plot of $m(\boldsymbol{x}_i)$ versus Y_i on the vertical axis. The identity line with zero intercept and unit slope is $E(Y_i) = m(\boldsymbol{x}_i)$. Let (L_i, U_i) be the asymptotically optimal population 95% prediction interval containing $m(\boldsymbol{x}_i)$. For example, if the errors are iid $N(0, \sigma^2)$, then $Y_i|m(\boldsymbol{x}_i) \sim N(m(\boldsymbol{x}_i), \sigma^2)$, and $(L_i, U_i) =$

$(m(\boldsymbol{x}_i) - 1.96\sigma, m(\boldsymbol{x}_i) + 1.96\sigma)$. Then the upper line has unit slope and passes through $(m(\boldsymbol{x}_i), U_i)$ while the lower line has unit slope and passes through $(m(\boldsymbol{x}_i), L_i)$.

The geometry of the "prediction region" for PI (9) is a natural sample analog of the population "asymptotically optimal prediction region". A response plot of $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$ versus Y_i has identity line $\hat{E}(Y_i) = \hat{m}(\boldsymbol{x}_i)$. The region corresponding to pointwise prediction intervals is between two lines with unit slope passing through the points $(\hat{m}(\boldsymbol{x}_i), \hat{U}_i)$ and $(\hat{m}(\boldsymbol{x}_i), \hat{L}_i)$, respectively, where $(\hat{L}_i, \hat{U}_i)$ is the asymptotically optimal prediction interval (9) for Y_f if $\boldsymbol{x}_f = \boldsymbol{x}_i$. For the multiple linear regression model, expect the points in the response plot to scatter in an evenly populated band for $n > 5p$. Other regression models, such as additive models, may need a much larger sample size n. See Section 3.1 for an example and simulations.

2.2 Prediction Regions

Asymptotically optimal prediction regions use ideas similar to those in the previous subsection. Some notation is needed. Let the ith case $\boldsymbol{x}_i$ be a $p \times 1$ random vector, and suppose the n cases are collected in an $n \times p$ matrix $\boldsymbol{X}$ with rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$.

The classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ of multivariate location and dispersion is the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \text{ and } \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})(\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})^\mathrm{T}. \tag{10}$$

Some important joint distributions for $\boldsymbol{x}$ are completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. An important model is the elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with probability density function $f(z) = k_p|\boldsymbol{\Sigma}|^{-1/2} g[(z - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(z - \boldsymbol{\mu})]$ where $k_p > 0$ is some constant and g is some known function. The multivariate normal (MVN) $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution is a special case.

Let the $p \times 1$ column vector $T(\boldsymbol{X})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{X})$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X})) = (\boldsymbol{x}_i - T(\boldsymbol{X}))^T \boldsymbol{C}^{-1}(\boldsymbol{X})(\boldsymbol{x}_i - T(\boldsymbol{X})) \tag{11}$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{X})$ is $D_i(T(\boldsymbol{X}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Often the data $\boldsymbol{X}$ will be suppressed. Then the classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. Following Johnson (1987, pp. 107-108), the population squared Mahalanobis distance

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{12}$$

and for elliptically contoured distributions, U has probability density function (pdf)

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{13}$$

The volume of the hyperellipsoid

$$\{z : (z - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1}(z - \overline{\boldsymbol{x}}) \leq h^2\} \text{ is equal to } \frac{2\pi^{\mathrm{p}/2}}{\mathrm{p}\Gamma(\mathrm{p}/2)} \mathrm{h}^\mathrm{p} \sqrt{det(\boldsymbol{S})}, \tag{14}$$

see Johnson and Wichern (1988, pp. 103-104).

Note that if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\, \boldsymbol{\Sigma})$, then

$$D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) = (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - d^{-1}\boldsymbol{\Sigma}^{-1} + d^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-1/2}).$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $d^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi_p^2$.

Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\, \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data. For $h > 0$, the hyperellipsoid

$$\{z : (z - T)^T \boldsymbol{C}^{-1}(z - T) \leq h^2\} = \{z : D_{\boldsymbol{z}}^2 \leq h^2\} = \{z : D_{\boldsymbol{z}} \leq h\} \tag{15}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)} \tag{16}$$

by (14). A future observation (random vector) $\boldsymbol{x}_f$ is in region (15) if $D_{\boldsymbol{x}_f} \leq h$.

A large sample $(1-\delta)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \overset{P}{\rightarrow} 1-\delta$. Let q_n be given by (6).

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then (15) is a large sample $(1-\delta)100\%$ prediction region if $h = D_{(up)}$ where $D_{(up)}$ is the q_nth sample quantile of the D_i. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, then region (15) is asymptotically optimal on a large class of elliptically contoured distributions in that its volume converges in probability to the volume of the minimum volume covering region $\{z : (z-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(z-\boldsymbol{\mu}) \leq u_{1-\delta}\}$ where $P(U \leq u_{1-\delta}) = 1-\delta$ and U has pdf given by (13). The classical parametric multivariate normal large sample prediction region uses $D_{\boldsymbol{x}_f}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \sqrt{\chi^2_{p,1-\delta}}$.

Notice that for the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, if $\boldsymbol{C}^{-1}$ exists, then $100q_n\%$ of the n cases are in the prediction region, and $q_n \rightarrow 1-\delta$ even if $(T, \boldsymbol{C})$ is not a good estimator. Hence the coverage q_n of the data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(T, \boldsymbol{C})$ is used or if the $\boldsymbol{x}_i$ do not come from an elliptically contoured distribution. Also notice that $q_n = 1-\delta/2$ or $q_n = 1-\delta+0.05$ for $n \leq 20p$ and $q_n \rightarrow 1-\delta$ as $n \rightarrow \infty$. If $q_n \equiv 1-\delta$, then (15) is a large sample prediction region, but taking q_n given by (6) improves the finite sample performance of the region. Taking $q_n \equiv 1-\delta$ does not take into account variability of $(T, \boldsymbol{C})$, and for small n the resulting prediction region tended to have undercoverage as high as $\min(0.05, \alpha/2)$. Using (6) helped reduce undercoverage for small n due to the unknown variability of $(T, \boldsymbol{C})$.

The Olive and Hawkins (2010) RMVN estimator $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ is an easily computed $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ under regularity conditions (E1) that include a large class of elliptically contoured distributions, and $c = 1$ for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Also see Zhang, Olive and Ye (2012). The RMVN estimator also gives a useful estimate of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data even when certain types of outliers are present.

Three new prediction regions will be considered. The nonparametric region uses the classical estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = D_{(up)}$. The semiparametric region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(up)}$. The parametric MVN region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{p,q_n}$ where $P(W \leq \chi^2_{p,q_n}) = q_n$ if $W \sim \chi^2_p$. All three regions are asymptotically optimal for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributions with nonsingular $\boldsymbol{\Sigma}$. The first two regions are asymptotically optimal for a large class of elliptically contoured distributions. For distributions with nonsingular covariance matrix $c_X\boldsymbol{\Sigma}$, the nonparametric region is a large sample $(1-\delta)100\%$ prediction region, but regions with smaller volume may exist. See Section 3.2 for examples and simulations.

3. Results

3.1 Regression

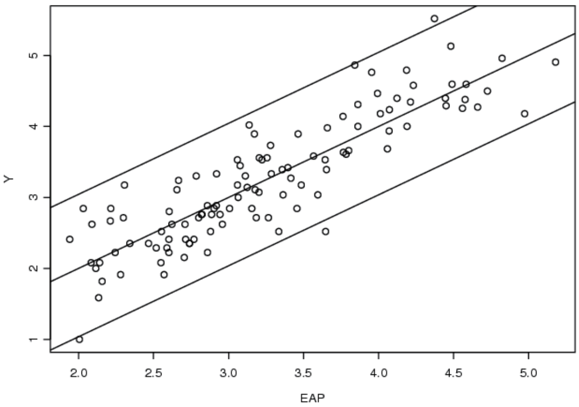


Figure 1. Pointwise prediction interval bands for Ozone data

Example 1 Chambers and Hastie (1993, pp. 251, 516) examine an environmental study that measured the four variables Y = *ozone concentration*, x_1 = *solar radiation*, x_2 = *temperature*, and x_3 = *wind speed* for $n = 111$

consecutive days. Figure 1 shows the response plot made in *Splus* with the pointwise large sample 95% PI bands for the additive model $Y = m(\boldsymbol{x}) + e$ where the additive predictor $m(\boldsymbol{x}) = \alpha + \sum_{j=1}^{3} S_j(x_j)$ for some functions S_j to be estimated. Here $\hat{m}(\boldsymbol{x})$ = estimated additive predictor (EAP). Note that the plotted points scatter about the identity line in a roughly evenly populated band, and that 3 of the 111 PIs (9) corresponding to the observed data do not contain Y.

A small simulation study compares the PI lengths and coverages for sample sizes $n = 50, 100$ and 1000 for PIs (8) and (9). Values for PI (8) were denoted by scov and slen while values for PI (9) were denoted by ocov and olen. The five error distributions in the simulation were 1) N(0,1), 2) t_3, 3) exponential(1) -1, 4) uniform$(-1, 1)$ and 5) $0.9N(0, 1) + 0.1N(0, 100)$. The value $n = \infty$ gives the asymptotic coverages and lengths and does not depend on the model. So these values are same for multiple linear and nonlinear regression as well as additive models.

Software for the simulations is described in Section 4. The multiple linear regression model with $E(Y_i) = 1 + x_{i1} + \cdots + x_{i7}$ was used. The vectors $(x_1, ..., x_7)^T$ were iid $N_7(\boldsymbol{0}, \boldsymbol{I}_7)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Another regression model was $Y_i = m(\boldsymbol{x}_i) + e_i$, $E(Y_i) = m(\boldsymbol{x}_i) = \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \beta_4 x_{i2}^2 + \beta_5 x_{i3} + \beta_6 x_{i3}^2$. This model was fit as an additive model in x_1, x_2, and x_3. The model was also fit with nonlinear regression where the mean function is known up to the six parameters, although then the second order multiple linear regression model is appropriate. For the additive model, the additive predictor $m(\boldsymbol{x}_i) = \alpha + \sum_{j=1}^{3} S_j(x_{ij})$. Both the nonlinear regression and additive model had the same mean function $m(\boldsymbol{x}_i) = x_{i1} + x_{i1}^2$. Thus $\boldsymbol{\beta} = (1, 1, 0, 0, 0, 0)^T$, $\alpha = 0$, $S_1(x_{i1}) = x_{i1} + x_{i1}^2$, $S_2(x_{i2}) = 0$ and $S_3(x_{i3}) = 0$. For these two models, the vectors $(x_1, x_2, x_3)^T$ were iid $N_3(\boldsymbol{0}, \boldsymbol{I}_3)$.

The Olive (2007) PIs (4) and (5) are tailored for multiple linear regression but are liberal (too short) for moderate n for many other techniques. The new PIs (8) and (9) are meant to have coverage near or higher than the nominal coverage for moderate n and for a wide variety of techniques and are longer than PIs (4) and (5). For multiple linear regression, the new PIs (8) and (9) were conservative (too long with roughly 98% coverage for the 95% PI and 70% or 60% coverage for the 50% PI) for $n = 50$ and 100 compared to (4) and (5) for least squares, least absolute deviations L_1 and an M-estimator using the *Splus* functions `l1fit` and `rreg`. See MathSoft (1999, pp. 293-295).

Table 1. PIs for additive models

error		95%	PI	95%	PI	50%	PI	50%	PI
type	n	slen	olen	scov	ocov	slen	olen	scov	ocov
1	50	5.126	4.998	0.959	0.950	1.862	1.674	0.596	0.520
1	100	4.691	4.515	0.968	0.957	1.662	1.528	0.570	0.516
1	1000	3.994	3.944	0.954	0.949	1.379	1.351	0.514	0.505
1	∞	3.920	3.920	0.95	0.950	1.349	1.349	0.50	0.50
2	50	9.444	8.630	0.951	0.943	2.385	2.153	0.576	0.512
2	100	8.245	7.596	0.962	0.954	2.042	1.878	0.577	0.532
2	1000	6.523	6.388	0.950	0.946	1.584	1.553	0.499	0.489
2	∞	6.365	6.365	0.950	0.950	1.530	1.530	0.50	0.50
3	50	5.186	4.823	0.958	0.948	1.573	1.275	0.611	0.526
3	100	4.677	4.156	0.967	0.955	1.382	1.063	0.603	0.533
3	1000	3.771	3.227	0.954	0.952	1.112	0.774	0.509	0.512
3	∞	3.664	2.996	0.950	0.950	1.099	0.693	0.50	0.50
4	50	2.634	2.598	0.961	0.958	1.237	1.087	0.593	0.506
4	100	2.318	2.272	0.972	0.968	1.155	1.028	0.561	0.480
4	1000	1.936	1.926	0.959	0.954	1.014	0.969	0.499	0.486
4	∞	1.900	1.900	0.950	0.950	1.00	1.00	0.50	0.50
5	50	19.689	17.747	0.944	0.935	2.976	2.693	0.608	0.548
5	100	18.754	16.230	0.955	0.946	2.352	2.164	0.580	0.534
5	1000	13.855	12.930	0.946	0.943	1.602	1.569	0.510	0.504
5	∞	13.490	13.490	0.950	0.950	1.507	1.507	0.50	0.50

The PIs (8) and (9) for nonlinear regression and additive models appear to have coverage near the nominal values in the simulations. For $n = 50$ and 100, the PIs for nonlinear regression were usually roughly 10% longer than those for additive models. The PIs for the additive model were computed using the *R* function `gam`. See Hastie and Tibshirani (1990) and Wood (2006). The PI (8) is not asymptotically optimal with error type 3. It is not known whether $\hat{m}$ is a consistent estimator of m, but the prediction intervals appear to have the correct asymptotic coverage and length. Some consistency results for the additive model and models of the form $Y = m(\boldsymbol{x}) + e$ where m is smooth are given in Müller, Schick and Wefelmeyer (2012) and Wang, Liu, Liang and Carroll (2011).

The simulation used 5000 runs and gave the proportion $\hat{p}$ of runs where Y_f fell within the nominal $100(1-\delta)\%$ PI. The count $m\hat{p}$ has a binomial$(m = 5000, p = 1 - \tau_n)$ distribution where $1 - \tau_n$ converges to the asymptotic coverage $(1 - \tau)$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/5000} = 0.0031$ and 0.0071 for $p = 0.05$ and 0.5, respectively. Hence an observed coverage $\hat{p} \in (.941, .959)$ for 95% and $\hat{p} \in (.479, .521)$ for 50% PIs suggests that there is no reason to doubt that the PI has the nominal coverage.

Table 1 shows that for $n = 1000$, the coverages and lengths are near the asymptotic $n = \infty$ values. For the 95% PI (9), the coverages were in or near (.94, .96) while the 50% PI (9) was sometimes slightly conservative. The coverage for the 50% PI (8) was near 60% for $n = 50$. PI (9) is recommended since its asymptotic optimality does not depend on the symmetry of the error distribution.

3.2 Prediction Regions

Rousseeuw and Van Driessen (1999) introduce the DD plot of the classical Mahalanobis distances MD versus the robust distances RD. Olive (2002) shows that if consistent estimators are used and n is large, then the plotted points will follow the identity line with unit slope and zero intercept if the data distribution is multivariate normal, and the plotted points will follow some other line through the origin if the data distribution is from a large class of elliptically contoured distributions but not multivariate normal.

Example 2 Buxton (1920) gives five measurements on 87 men: *height*, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. The 5 outliers have *heights* that were recorded to be about 19mm and head lengths recorded as the heights. The DD plot of the classical Mahalanobis distances MD versus the RMVN distances RD can be used to visualize the prediction regions. Figure 2 shows the DD plot where points to the left of the vertical line are in the nonparametric large sample 90% prediction region. Points below the horizontal line are in the semiparametric region. The horizontal line at $RD = 3.33$ corresponding to the parametric MVN 90% region is obscured by the identity line. This region contains 78 of the cases. Since $n = 87$, the nonparametric and semiparametric regions used the 95th quantile. Since there were 5 outliers, this quantile was a linear combination of the largest clean distance and the smallest outlier distance. The semiparametric 90% region blows up unless the outlier proportion is small.

Figure 3 shows the DD plot and 3 prediction regions after the 5 outliers were removed. The classical and robust distances cluster about the identity line and the three regions are similar, with the parametric MVN region cutoff again at 3.33, slightly below the semiparametric region cutoff of 3.44.

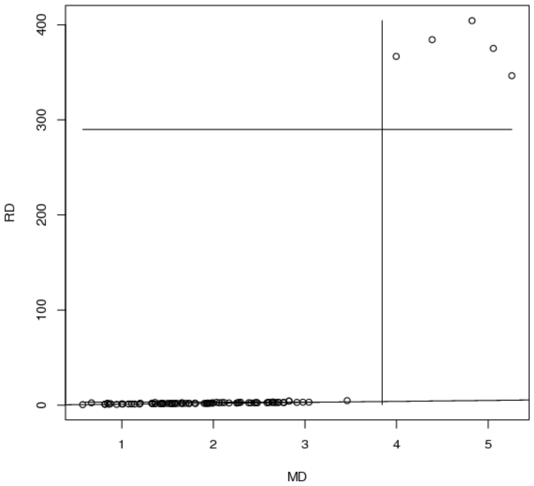


Figure 2. Prediction regions for Buxton data

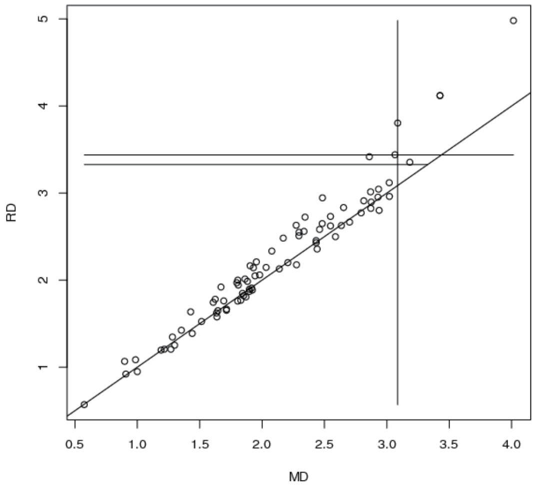


Figure 3. Prediction regions for Buxton data without outliers

Example 3 Cook and Weisberg (1999, pp. 351, 433, 447) give a data set on 82 mussels sampled off the coast of New Zealand. The variables are $X_1 = \log(S)$, $X_2 = \log(M)$, $X_3 = L, X_4 = \log(W)$, and $X_5 = height$ where S is the *shell mass*, M is the *muscle mass* in grams, L is the *length L*, W is the *shell width* and H is the *height* of the shell in mm. Figure 4 shows a DD plot of the data with multivariate prediction regions added. This plot suggests that the data may come from an elliptically contoured distribution that is not multivariate normal. The semiparametric and nonparametric 90% prediction regions consist of the cases below the $RD = 5.86$ line and to the left of the $MD = 4.41$ line. These two lines intersect on a line through the origin that is followed by the plotted points. The parametric MVN prediction region is given by the points below the $RD = 3.33$ line and does not contain enough cases. Points to the left of a vertical line $MD = 3.33$ would give a modified classical MVN prediction region. Parametric prediction regions for multivariate normal data tend to have severe undercoverage if the data is not multivariate normal. This undercoverage problem becomes worse as p increases, since if the cutoff h is too small, then the volume of the prediction region depends on h^p by (14).

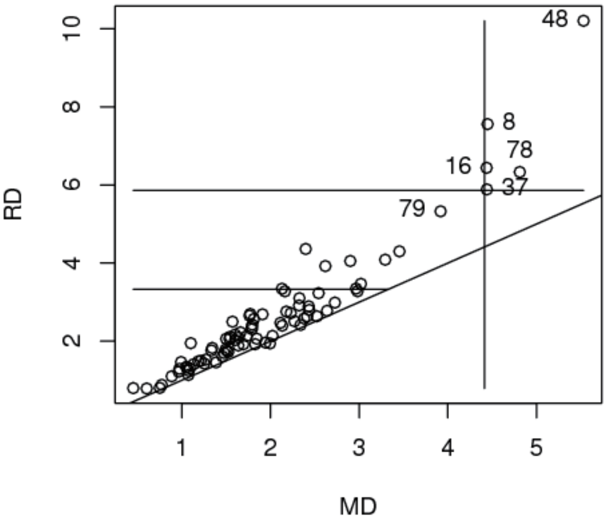


Figure 4. DD plot of the Mussels data

Simulations for the prediction regions used $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ where $\boldsymbol{A} = diag(\sqrt{1}, \sqrt{2}, ..., \sqrt{p})$, $\boldsymbol{w} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{w} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{w} \sim MVT_p(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy(0,1). All simulations used 5000 runs and $\delta = 0.1$.

Table 2. Coverages for 90% Prediction Regions

$\boldsymbol{w}$ dist	n	p	ncov	scov	mcov	voln	volm
MVN	600	30	0.906	0.919	0.902	0.503	0.512
MVN	1500	30	0.899	0.899	0.900	1.014	1.027
LN	1000	10	0.903	0.906	0.567	0.659	0+
MVT(1)	1000	10	0.914	0.914	0.541	22634.3	0+

For large n, the semiparametric and nonparametric regions are likely to have coverage near 0.90 because the coverage on the training sample is slightly larger than 0.9 and $\boldsymbol{x}_f$ comes from the same distribution as the $\boldsymbol{x}_i$. For $n = 10p$ and $2 \leq p \leq 40$, the semiparametric region had coverage near 0.9. The ratio of the volumes

$$\frac{h_i^p \sqrt{det(\boldsymbol{C}_i)}}{h_2^p \sqrt{det(\boldsymbol{C}_2)}}$$

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric region, and $i = 3$ was the parametric MVN region. The volume ratio converges in probability to 1 for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data, and the ratio converges to 1 for $i = 1$ on a large class of elliptically contoured distributions. The parametric MVN region often had coverage much lower than 0.9 with a volume ratio near 0, recorded as 0+. The volume ratio tends to be tiny when the coverage is much less than the nominal value 0.9. For $10p \leq n \leq 20p$, the nonparametric region often had good coverage and volume ratio near 0.5.

Simulations and Table 2 suggest that for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data, the coverages (ncov, scov and mcov) for the 3 regions are near 90% for $n = 20p$ and that the volume ratios voln and volm are near 1 for $n = 50p$. With fewer than 5000 runs, this result held for $2 \leq p \leq 80$. For the non-elliptically contoured LN data, the nonparametric region had voln well under 1, but the volume ratio blew up for $\boldsymbol{w} \sim MVT_p(1)$.

4. Discussion

4.1 General Comments

There are not many practical competitors for the new prediction intervals and regions. Parametric prediction intervals and regions usually assume normality and tend to have severe undercoverage when the normality assumption does not hold. For confidence intervals and testing, misspecification of normality is sometimes not too important if the estimators are asymptotically normal, but for parametric prediction intervals and regions, correct specification of the parametric model is important. For example, do not use a parametric prediction region based on the multivariate normal distribution if the plotted points in the DD plot fail to cover the identity line.

Another competitor for regression is bootstrap prediction intervals. These PIs take hundreds of times longer to compute than PI (9), and convergence problems are greatly multiplied for models such as nonlinear regression models. Also bootstrap PIs may not be valid if a fixed number B of bootstrap samples are used. Di Bucchianico, Einmahl and Mushkudiani (2001) use the minimum volume ellipsoid (MVE) estimator to cover m out of n cases to produce MVE tolerance regions, but the technique can only be used on tiny data sets.

The location model is a special case of both the regression model (1) and of the multivariate location and dispersion model. Let $a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+1}{n-1}}$. Let $c = \lceil n(1-\delta) \rceil$. Let shorth$(c) = (Y_{(d)}, Y_{(d+c-1)})$. Let MED$(n)$ be the sample median. If $Y_1, ..., Y_n$ are iid, then the recommended large sample $100(1-\delta)$% PI for Y_f is the closed interval $[L_n, U_n] = [(1-a_n)\text{MED}(n) + a_n Y_{(d)}, (1-a_n)\text{MED}(n) + a_n Y_{(d+c-1)}]$. This PI is (5) using the least absolute deviations estimator, but with a closed interval.

Simulations were done in *Splus* and *R*. See R Development Core Team (2008). The Buxton data and programs in the collection of functions *rpack.txt* are available at (www.math.siu.edu/olive/ol-bookp.htm). For multiple linear regression, the function `pisim` simulates PIs (4) and (5) while the *Splus* function `pisim4` simulates PIs (8) and (9) using OLS, L_1 and M-estimators. The function `pisim3` was used to create Table 1 while `pisim5` uses `nls` to simulate PIs for nonlinear regression. Care is needed when using `pisim5` since for some versions of *R/Splus*, the `nls` function will fail to converge for some runs. Using nruns = 500 is less likely to cause an error than nruns=5000. The function *predsim* was used for Table 2. The function *ddplot4* was used to produce Figures 2, 3 and 4. The function *lpisim* simulates the PI for the location model while *covrmvn* computes the RMVN estimator.

4.2 Conclusions

Parametric prediction intervals and regions are notorious for severe undercoverage. The new techniques are designed to have good coverage at the training data, even if the model assumptions fail to hold. The Olive (2007) PIs (4) and (5) are tailored for multiple linear regression but are too short for many other techniques for moderate n. PIs (8) and (9) are generally longer than PIs (4) and (5) and have coverage near or higher than the nominal value for many techniques even for moderate n, say $n > 10$ (model degrees of freedom). PIs (8) and (9) are quite conservative for multiple linear regression for moderate n. These PIs are useful since the error distribution does not need to be known.

The new nonparametric and semiparametric prediction regions appear to have good coverage for $n > 20p$ and may be the first easily computed prediction regions that are effective when the underlying multivariate distribution is unknown.

For the prediction regions, use the DD plot to check the multivariate normality assumption and to check for the presence of outliers. If $n > 20p$ and the plotted points cluster tightly about a line through the origin, then the nonparametric and semiparametric prediction regions may have good coverage. For regression with additive errors, if n is large and the plotted points cluster about the identity line in the response plot, then the new prediction intervals may have good coverage.

Acknowledgements

The author thanks the editor and two referees for their comments that improved this article.

References

Buxton, L. H. D. (1920). The anthropology of Cyprus. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland, 50*, 183-235. http://dx.doi.org/10.2307/2843379

Cai, T., Tian, L., Solomon, S. D., & Wei, L. J. (2008). Predicting future responses based on possibly misspecified working models. *Biometrika, 95*, 75-92. http://dx.doi.org/10.1093/biomet/asm078

Chambers, J. M., & Hastie, T. J. (Eds.) (1993). *Statistical models in S*. New York, NY: Chapman & Hall.

Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics.* New York, NY: John Wiley & Sons.

Di Bucchianico, A., Einmahl, J. H. J., & Mushkudiani, N. A. (2001). Smallest nonparametric tolerance regions. *The Annals of Statistics, 29*, 1320-1343. http://dx.doi.org/10.1214/aos/1013203456

Grübel, R. (1988). The length of the shorth. *The Annals of Statistics, 16*, 619-628. http://dx.doi.org/10.1214/aos/1176350823

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London, UK: Chapman & Hall.

Johnson, M. E. (1987). *Multivariate statistical simulation.* New York, NY: John Wiley & Sons.

Johnson, R. A., & Wichern, D. W. (1988). *Applied multivariate statistical analysis* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

MathSoft. (1999). *S-plus 2000 guide to statistics, Vol. 1.* Seattle, WA: MathSoft.

Müller, U. U., Schick, A., & and Wefelmeyer, W. (2012). Estimating the error distribution function in semiparametric additive regression models. *Journal of Statistical Planning and Inference, 142*, 552-566. http://dx.doi.org/10.1016/j.jspi.2011.08.013

Olive, D. J. (2002). Applications of robust distances for regression. *Technometrics, 44,* 64-71. http://dx.doi.org/10.1198/004017002753398335

Olive, D. J. (2007). Prediction intervals for regression models. *Computational Statistics and Data Analysis, 51*, 3115-3122. http://dx.doi.org/10.1016/j.csda.2006.02.006

Olive, D. J., & Hawkins, D. M. (2003). Robust regression with high coverage. *Statistics and Probability Letters, 63*, 259-266. http://dx.doi.org/10.1016/S0167-7152(03)00090-7

Olive, D. J., & Hawkins, D. M. (2010). Robust multivariate location and dispersion. Retrieved from http://www.math.siu.edu/olive/preprints.htm

R Development Core Team. (2008). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, NY: John Wiley & Sons.

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*, 212-223. http://dx.doi.org/10.1080/00401706.1999.10485670

Wang, L., Liu, X., Liang, H., & Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics, 39,* 1827-1851. http://dx.doi.org/10.1214/11-AOS885SUPP

Welsh, A. H. (1986). Bahadur representation for robust scale estimators based on regression residuals. *The Annals of Statistics, 14,* 1246-1251. http://dx.doi.org/10.1214/aos/1176350064

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Boca Rotan, FL: Chapman & Hall/CRC.

Zhang, J., Olive, D. J., & Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability, 1*, 119-136. http://dx.doi.org/10.5539/ijsp.v1n2p119

International Journal of Statistics and Probability; Vol. 2, No. 1; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The RS Generalized Lambda Distribution Based Calibration Model

Steve Su[1], Abeer Hasan[2] & Wei Ning[2]

[1] Covance Pty Ltd., Sydney, Australia; School of Mathematics and Statistics, University of Western Australia, Crawley, Australia

[2] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, USA

Correspondence: Steve Su, Covance Pty Ltd., Sydney, Australia; School of Mathematics and Statistics, University of Western Australia, Crawley 6009, Australia. E-mail: dbarro2@gmail.com

Received: December 5, 2012 Accepted: January 10, 2013 Online Published: January 22, 2013

doi:10.5539/ijsp.v2n1p101 URL: http://dx.doi.org/10.5539/ijsp.v2n1p101

Abstract

We propose a flexible linear calibration model with errors from RS (Ramberg & Schmeiser, 1974) generalized lambda distribution ($G\lambda D$). We demonstrate the derivation of the maximum likelihood estimates of RS $G\lambda D$ parameters and examine the estimation performance using a simulation study for sample sizes ranging from 30 to 200. The use of RS $G\lambda D$ calibration model not only provides statistical modeller with a richer range of distributional shapes, but can also provide more precise parameter estimates compared to the standard Normal calibration model or skewed Normal calibration model proposed by Figueiredoa, Bolfarinea, Sandovala and Limab (2010).

Keywords: generalized lambda distribution, linear calibration model, skew normal distribution, maximum likelihood estimation

1. Introduction

The statistical calibration model is a reverse regression technique, where we use the response variable to predict the corresponding explanatory variable. There are number of applications of this technique in science. For example, we may use radiometric dating to ascertain the age of a tree and further verify our result using tree rings. Our aim, however, is to use radiometric dating to estimate age of new trees, and the problem is whether we should minimize errors in the observation or minimize errors in age determination. There are many similar problems in substance concentration determination in biology and chemistry, physical quantities determination in physics and blood pressure/cholsterol level measurement in medicine. The literature on calibration problem has a long history, and one of the earliest works can be found in Eisenhart (1939).

The usual calibration experiment is a two stage process involving two random variables X and Y. The first stage is known as the calibration trial, where we observe the n values of the response variable $y_1, \cdots, y_n$ from a given set of explanatory values $x_1, \cdots, x_n$ and we can estimate the link function between X and Y. The second stage is known as the calibration experiment, where we observe $k \geq 1$ value(s) of the response variable Y as $y_{01}, \cdots, y_{0k}$ which are mapped from some unknown value x_0 from the explanatory variable X. We can express these two stages by the following equations.

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \cdots, n;$$
$$y_{0j} = \alpha + \beta x_0 + \varepsilon_{0j}, \quad j = 1, \cdots, k, \tag{1.1}$$

We usually assume that the errors $\varepsilon_1, \cdots, \varepsilon_n, \varepsilon_{01}, \cdots, \varepsilon_{0k}$ are i.i.d and Normally distributed with mean 0 and variance σ^2. Also, $x_1, \cdots, x_n$ are known and α, β, x_0 and σ^2 are unknown parameters which we need to estimate.

As an extension to Normal distribution, Azzalini (1985) introduced the skewed Normal distribution. The skewed Normal distribution is defined as

$$g(x; \xi, \omega, \lambda) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\lambda\left(\frac{x-\xi}{\omega}\right)\right), \tag{1.2}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the p.d.f. and c.d.f. of a standard normal distribution respectively. Specially, when $\xi = 0$ and $\omega = 1$, we obtain the standard skewed Normal distribution.

Based on (1.2), Figueiredoa et al. (2010) defined a skew-normal calibration model by assuming that ε_i and ε_{0j} are i.i.d. and follow a skewed Normal distribution with $\xi = 0$ denoted by $SN(0, \omega, \lambda)$. This gives us the following calibration model:

$$y_i|x_i \sim SN(\alpha + \beta x_i; \omega; \lambda), \quad i = 1. \cdots, n,$$

$$y_{0j}|x_0 \sim SN(\alpha + \beta x_i; \omega; \lambda), \quad j = 1, \cdots, k. \tag{1.3}$$

In (1.3), the conditional distribution of y_i given x_i and y_{0j} given x_0 are governed by skewed Normal distributions. This skewed Normal calibration model allows the modeller to cope with some degree of skewness in the error distribution. However, this is still limited as the skewed Normal distribution have limited range of shapes. The skewed Normal distribution still cannot handle heavy tailed, U shape, uniform, triangular or exponential upward/downward patterns. These shapes however, can be captured using $G\lambda D$ (generalized lambda distributions), and we propose a further extension to the calibration model by using RS $G\lambda D$.

Our article is organized as follows. In Section 2, we introduce the $G\lambda D$ family. In Section 3, we outline the RS $G\lambda D$ calibration model and discuss possible ways to estimate parameters of the model using maximum likelihood estimation. In Section 4, we demonstrate the estimation performance of our proposed model across a range of different sample sizes from 30 to 200. As a further test to our proposed model to the literature, we compare the performance of RS $G\lambda D$ calibration model against Normal and skewed Normal calibration model with respect to a real life dataset used by Figueiredoa et al. (2010) in Section 5. A discussion of our proposed method is given in Section 6.

2. Generalized Lambda Distributions

The RS $G\lambda D$ (Ramberg & Schmeiser, 1974) is an extension of Tukey's lambda distribution. It is defined by its inverse distribution function:

$$F^{-1}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2} \qquad 0 \leq u \leq 1 \tag{2.1}$$

From (2.1), $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are respectively the location, inverse scale and shape 1 and shape 2 parameters. Karian and Dudewicz (2000) noted that $G\lambda D$ is defined only if $\frac{\lambda_2}{\lambda_3 u^{\lambda_3 - 1} + \lambda_4 (1-u)^{\lambda_4 - 1}} \geq 0$ for $0 \leq u \leq 1$. The conditions for which RS $G\lambda D$ is a valid p.d.f. are set out in Karian and Dudewicz (2000) and these are also programmed in GLDEX package in R (Su, 2010, 2007a).

Freimer, Kollia, Mudholkar and Lin (1988) describe another distribution known as FKML $G\lambda D$. The FKML $G\lambda D$ can be written as:

$$F^{-1}(u) = \lambda_1 + \frac{\frac{u^{\lambda_3}-1}{\lambda_3} - \frac{(1-u)^{\lambda_4}-1}{\lambda_4}}{\lambda_2} \qquad 0 \leq u \leq 1 \tag{2.2}$$

Under (2.2), $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are respectively the location, inverse scale and shape 1 and shape 2 parameters.

The fundamental motivation for the development of FKML $G\lambda D$ is that the distribution is defined over all λ_3 and λ_4 (Freimer et al., 1988). The only restriction on FKML $G\lambda D$ is $\lambda_2 > 0$. This is more convenient to deal with computationally than RS $G\lambda D$ and hence it is sometimes the preferred $G\lambda D$ for some researchers.

We restrict our attention in this article to the more difficult problem of fitting RS $G\lambda D$ calibration model to data. Without loss of generality, the method we outlined below can be easily adapted to build FKML $G\lambda D$ calibration model.

3. Statistical Model

3.1 GλD Based Calibration Model

We consider the following usual calibration model:

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \cdots, n, \tag{3.1}$$

$$y_{0j} = \alpha + \beta x_0 + \epsilon_j, \quad j = 1, \cdots, k. \tag{3.2}$$

We assume that ϵ_i and ϵ_j are i.i.d. $G\lambda D(0, \lambda_2, \lambda_3, \lambda_4)$. In general, we consider $x_1, \cdots, x_n$ to be known and fixed and α,β, λ_2, λ_3 and λ_4 are parameters we need to estimate. Our $G\lambda D$ calibration model takes the following form:

$$y_i|x_i \sim G\lambda D(\alpha + \beta x_i, \lambda_2, \lambda_3, \lambda_4), \tag{3.3}$$

$$y_{0j}|x_0 \sim G\lambda D(\alpha+\beta x_0, \lambda_2, \lambda_3, \lambda_4). \tag{3.4}$$

Consequently, the likelihood function for RS $G\lambda D$ is:

$$L(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{y}_0) = \prod_{i=1}^{n} \frac{\lambda_2}{\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_i)^{\lambda_4-1}} \cdot \prod_{j=1}^{k} \frac{\lambda_2}{\lambda_3 z_j^{\lambda_3-1} + \lambda_4(1-z_j)^{\lambda_4-1}}, \tag{3.5}$$

where

$$y_i = (\alpha+\beta x_i) + \frac{z_i^{\lambda_3} - (1-z_i)^{\lambda_4}}{\lambda_2},$$
$$y_{0j} = (\alpha+\beta x_0) + \frac{z_j^{\lambda_3} - (1-z_j)^{\lambda_4}}{\lambda_2},$$

and $0 \le z_i, z_j \le 1, \boldsymbol{\theta} = (\alpha,\beta, x_0, \lambda_2, \lambda_3, \lambda_4)$.

3.2 Estimation of Parameters

From (3.5), we obtain the following log likelihood function:

$$\log L(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{y_0}) = \sum_{i=1}^{n} \log\left(f_1(\boldsymbol{\theta}, y_i)\right) + \sum_{j=1}^{k} \log\left(f_2(\boldsymbol{\theta}, y_{0j})\right) \tag{3.6}$$

where

$$f_1(\boldsymbol{\theta}, y_i) = \frac{\lambda_2}{\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_i)^{\lambda_4-1}},$$
$$f_2(\boldsymbol{\theta}, y_{0j}) = \frac{\lambda_2}{\lambda_3 z_j^{\lambda_3-1} + \lambda_4(1-z_j)^{\lambda_4-1}}$$

Taking the derivative of (3.6), we obtain the following:

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{1}{f_1}\frac{\partial f_1}{\partial \boldsymbol{\theta}} + \sum_{j=1}^{k} \frac{1}{f_2}\frac{\partial f_2}{\partial \boldsymbol{\theta}}, \tag{3.7}$$

where $\boldsymbol{\theta} = (\alpha,\beta, x_0, \lambda_2, \lambda_3, \lambda_4)$.

Theoretically, the MLE of $\boldsymbol{\theta}$ is the solution of (3.7) when it is set to be equal to 0. The derivatives $\frac{\partial f_1}{\partial \boldsymbol{\theta}}$ and $\frac{\partial f_2}{\partial \boldsymbol{\theta}}$ are given below.

$$\begin{aligned}
\frac{\partial f_1}{\partial \lambda_2} &= \frac{\partial f_1}{\partial z_i} \cdot \frac{\partial z_i}{\partial y_i} \cdot \frac{\partial y_i}{\partial \lambda_2} \\
&= \left(\lambda_2 \frac{-\lambda_3(\lambda_3-1)z_i^{\lambda_3-2} + \lambda_4(\lambda_4-1)(1-z_i)^{\lambda_4-2}}{(\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_i)^{\lambda_4-1})^2}\right) \cdot \left(\frac{\lambda_2}{\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_i)^{\lambda_4-1}}\right) \cdot \left(-\frac{z_i^{\lambda_3} - (1-z_i)^{\lambda_4}}{\lambda_2^2}\right) \\
&= \frac{[\lambda_3(\lambda_3-1)z_i^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_i)^{\lambda_4-2}](z_i^{\lambda_3} - (1-z_i)^{\lambda_4})}{(\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_i)^{\lambda_4-1})^3} \\
\frac{\partial f_1}{\partial \lambda_3} &= (-\lambda_2)\frac{[\lambda_3(\lambda_3-1)z_i^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_i)^{\lambda_4-2}](z_i^{\lambda_3}\log z_i)}{(\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_i)^{\lambda_4-1})^3} \\
\frac{\partial f_1}{\partial \lambda_4} &= \lambda_2\frac{[\lambda_3(\lambda_3-1)z_i^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_i)^{\lambda_4-2}]((1-z_i)^{\lambda_3}\log(1-z_i))}{(\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_i)^{\lambda_4-1})^3} \\
\frac{\partial f_1}{\partial \alpha} &= (-\lambda_2^2)\frac{[\lambda_3(\lambda_3-1)z_i^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_i)^{\lambda_4-2}]}{(\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_i)^{\lambda_4-1})^3} \\
\frac{\partial f_1}{\partial \beta} &= (-\lambda_2^2)\frac{[\lambda_3(\lambda_3-1)z_i^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_i)^{\lambda_4-2}] \cdot x_i}{(\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_i)^{\lambda_4-1})^3} \\
\frac{\partial f_2}{\partial \lambda_2} &= \frac{[\lambda_3(\lambda_3-1)z_j^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_j)^{\lambda_4-2}](z_j^{\lambda_3} - (1-z_j)^{\lambda_4})}{(\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_j)^{\lambda_4-1})^3}
\end{aligned}$$

$$\frac{\partial f_2}{\partial \lambda_3} = (-\lambda_2)\frac{[\lambda_3(\lambda_3-1)z_j^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_j)^{\lambda_4-2}](z_j^{\lambda_3}\log z_j)}{(\lambda_3 z_i^{\lambda_3-1} + \lambda_4(1-z_i)^{\lambda_4-1})^3}$$

$$\frac{\partial f_1}{\partial \lambda_4} = \lambda_2\frac{[\lambda_3(\lambda_3-1)z_j^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_j)^{\lambda_4-2}]((1-z_j)^{\lambda_3}\log(1-z_j))}{(\lambda_3 z_j^{\lambda_3-1} + \lambda_4(1-z_j)^{\lambda_4-1})^3}$$

$$\frac{\partial f_2}{\partial \alpha} = (-\lambda_2^2)\frac{[\lambda_3(\lambda_3-1)z_j^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_j)^{\lambda_4-2}]}{(\lambda_3 z_j^{\lambda_3-1} + \lambda_4(1-z_j)^{\lambda_4-1})^3}$$

$$\frac{\partial f_2}{\partial \beta} = (-\lambda_2^2)\frac{[\lambda_3(\lambda_3-1)z_j^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_j)^{\lambda_4-2}]\cdot x_0}{(\lambda_3 z_j^{\lambda_3-1} + \lambda_4(1-z_j)^{\lambda_4-1})^3}$$

$$\frac{\partial f_2}{\partial x_0} = (-\lambda_2^2)\frac{[\lambda_3(\lambda_3-1)z_j^{\lambda_3-2} - \lambda_4(\lambda_4-1)(1-z_j)^{\lambda_4-2}]\cdot \beta}{(\lambda_3 z_j^{\lambda_3-1} + \lambda_4(1-z_j)^{\lambda_4-1})^3}$$

It is difficult to obtain the exact solutions of setting (3.7) to zero using the above formulations, owing to the fact that RS $G\lambda D$ is defined by its inverse quantile function and there is a high degree of complexity involved in solving the above equations. As an alternative, we carry out the maximum likelihood estimation by maximising (3.6) directly using Nelder-Mead optimisation algorithm as is customary done for maximum likelihood estimation problems involving $G\lambda D$ (see Su, 2010, 2007a, 2007b). This is a preferred and more reliable method of estimation as opposed to trying to satisfy the exact conditions to which all of the above equations equal to zero. The GLDEX package in R (Su, 2010, 2007a) facilitates the Nelder-Mead optimisation algorithm for $G\lambda D$.

Our algorithm is as follows:

1) Generate a set of initial values for $\alpha, \beta, x_0, \lambda_2, \lambda_3, \lambda_4$. There are a number of strategies that can be used to determine the best set of initial values. One strategy is to generate initial values α, β, x_0 using Normal or skewed Normal calibration model and then generate some low discrepancy quasi random numbers for $\lambda_2, \lambda_3, \lambda_4$ over a range of values and select the set of initial values that maximises (3.6). Alternatively all initial values can be randomly generated using low discrepancy quasi random numbers.

2) Set $\lambda_1 = \alpha + \beta x_0$.

3) Check that $G\lambda D(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ is a valid statistical distribution, this can be done using GLDEX package in R.

4) Check the minimal support of $G\lambda D(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ is lower or equal to the lowest value of y_0. Similarly, check that the maximum support of $G\lambda D(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ is greater or equal to the largest value of y_0. This is to ensure that the fitted $G\lambda D$ will span the entire dataset. If these conditons are not met, choose another set of initial values and repeat from 2).

5) Conduct Nelder Mead optimisation by maximising (3.6) directly using the above initial values to obtain the required estimates.

4. Simulations

We conduct simulations to illustrate the performance of our RS $G\lambda D$ calibration model for sample size $n = 30, 50, 100$ and 200 with $\alpha = 3, \beta = 1.5, x_0 = 15$ or $40, \lambda_3 = 10, \lambda_4 = 1$, and $\lambda_2 = 2, 5, 10$. We further generate $x_1, x_2, \cdots, x_n$ from $Uniform(10, 30)$, and we set $k = 1$. We use the true parameters as our initial values to kick start the optimisation process to obtain our MLE estimate for x_0.

We repeat this process 1000 times, which give us 1000 $\hat{x}_{0m}$ estimates of x_0. The mean $\hat{x}_0$, Bias(x_0) and MSE(x_0) are calculated as follows:

$$\bar{\hat{x}}_0 = \frac{1}{1000}\sum_{m=1}^{1000}\hat{x}_{0m}$$

$$\text{Bias}(x_0) = \frac{1}{1000}\sum_{m=1}^{1000}(\hat{x}_{0m} - x_0)$$

$$\text{MSE}(x_0) = \frac{1}{1000}\sum_{m=1}^{1000}(\hat{x}_{0m} - x_0)^2$$

The results of above simulations are shown in Tables 1 and 2. As expected, the MSE decreases as we increase the sample size or increase the value of inverse scale parameter λ_2. In terms of bias, we observe that the performance appear to be fairly consistent across sample sizes, this gives confidence in the use of RS $G\lambda D$ calibration model for smaller samples, even though there are are more parameters that need to be estimated from this model. There also appears to be a tendency for RS $G\lambda D$ calibration model to slightly overestimate as nearly all the bias results are positive. Increasing the shape parameter λ_3 does not always result in increase in MSE, this is because the shape parameter spaces of λ_3 and λ_4 for RS $G\lambda D$ are fairly complex.

Table 1. Simulations results with $x_0 = 15, \alpha = 3, \beta = 1.5, \lambda_4 = 1$

			$\lambda_2 = 2$			$\lambda_2 = 5$			$\lambda_2 = 10$	
n	λ_3	$\hat{x}_0$	Bias	MSE	$\hat{x}_0$	Bias	MSE	$\hat{x}_0$	Bias	MSE
30	10	15.1105	0.1105	0.0263	15.0386	0.0386	0.0042	15.0178	0.0178	0.0010
50	10	15.0944	0.0944	0.0232	15.0352	0.0352	0.0040	15.0172	0.0172	0.0010
100	10	15.0994	0.0994	0.0184	15.0396	0.0396	0.0035	15.0185	0.0185	0.0008
200	10	15.1053	0.1053	0.0166	15.0340	0.0340	0.0030	15.0173	0.0173	0.0007
30	5	15.1430	0.1430	0.0292	15.0578	0.0578	0.0056	15.0285	0.0285	0.0012
50	5	15.1445	0.1445	0.0270	15.0530	0.0530	0.0047	15.0292	0.0292	0.0012
100	5	15.1381	0.1381	0.0214	15.0531	0.0531	0.0043	15.0264	0.0264	0.0010
200	5	15.1429	0.1429	0.0187	15.0534	0.0534	0.0038	15.0227	0.0227	0.0009
30	1	15.0271	0.0271	0.0244	15.0014	0.0014	0.0061	15.0040	0.0040	0.0017
50	1	15.0367	0.0367	0.0169	15.0030	0.0030	0.0048	14.9993	-0.0007	0.0014
100	1	15.0292	0.0292	0.0084	15.0093	0.0093	0.0030	15.0029	0.0029	0.0010
200	1	15.0262	0.0262	0.0052	15.0130	0.0130	0.0016	15.0022	0.0022	0.0007

Table 2. Simulations results with $x_0 = 40, \alpha = 3, \beta = 1.5, \lambda_4 = 1$

			$\lambda_2 = 2$			$\lambda_2 = 5$			$\lambda_2 = 10$	
n	λ_3	$\hat{x}_0$	Bias	MSE	$\hat{x}_0$	Bias	MSE	$\hat{x}_0$	Bias	MSE
30	10	40.1070	0.1070	0.0259	40.0375	0.0375	0.0049	40.0189	0.0189	0.0012
50	10	40.1051	0.1051	0.0235	40.0388	0.0388	0.0039	40.0177	0.0177	0.0009
100	10	40.1077	0.1077	0.0205	40.0353	0.0353	0.0031	40.0188	0.0188	0.0008
200	10	40.1088	0.1088	0.0169	40.0387	0.0387	0.0028	40.0184	0.0184	0.0008
30	5	40.1339	0.1339	0.0319	40.0557	0.0557	0.0064	40.0288	0.0288	0.0014
50	5	40.1391	0.1391	0.0302	40.0554	0.0554	0.0046	40.0280	0.0280	0.0013
100	5	40.1405	0.1405	0.0232	40.0479	0.0479	0.0039	40.0264	0.0264	0.0010
200	5	40.1538	0.1538	0.0236	40.0474	0.0474	0.0035	40.0205	0.0205	0.0007
30	1	40.0331	0.0331	0.0290	39.9984	-0.0016	0.0058	40.0035	0.0035	0.0016
50	1	40.0348	0.0348	0.0159	40.0031	0.0031	0.0045	40.0022	0.0022	0.0013
100	1	40.0311	0.0311	0.0099	40.0078	0.0078	0.0024	39.9996	-0.0004	0.0009
200	1	40.0217	0.0217	0.0036	40.0114	0.0114	0.0017	40.0031	0.0031	0.0007

Table 3. Simulations results with $x_0 = 15, \alpha = 3, \beta = 1.5$, true error distribution GEV(0.1860, 0.4016, 0.1511) is approximated by RS $G\lambda D$ with $\lambda_1 = 0, \lambda_2 \approx -0.0374, \lambda_3 \approx -0.0027, \lambda_4 \approx -0.0212$

n	$\hat{x}_0$	Bias	MSE
30	15.3140	0.3140	0.2149
50	15.3269	0.3269	0.2154
100	15.2815	0.2815	0.1774
200	15.2860	0.2860	0.1689

We further considered using RS $G\lambda D$ to approximate generalized extreme value distribution (GEV) with location, scale and shape parameters being 0.1860, 0.4016, 0.1511 respectively. We choose RS $G\lambda D$ with $\lambda_1 = 0, \lambda_2 \approx -0.0374, \lambda_3 \approx -0.0027, \lambda_4 \approx -0.0212$ for this demonstration (Figure 1). We then generate simulated data based on GEV and use our approximated RS $G\lambda D$ to estimate x_0 with $\alpha = 3, \beta = 1.5$ and repeat this over 1000 simulation runs. The result of this simulation is given in Table 3. We observe that the RS $G\lambda D$ calibration model tends to overestimate the true x_0 by a small margin, but the bias appears to decrease as sample size increases.

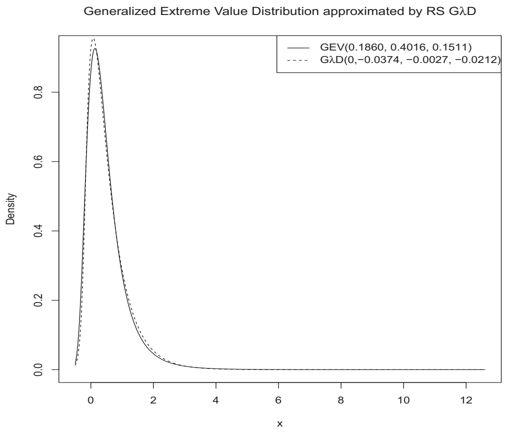


Figure 1. Approximating *GEV* using RS *GλD*

5. Application

We apply the RS *GλD* calibration model to a dataset which measures teenager testicular volume (ml^3). This dataset is from Chipkevitch, Nishimura, Tu and Galea-Rajas (1996) and consists of 42 observations. Figueiredoa et al. (2010) considered two measurement methods from Chipkevitch et al. (1996): dimensional measurement with a caliper (DM) and measurement by ultrasonography (US) and the data is given in Table 4. In their paper, Figueiredoa et al. (2010) consider the x_0 value of 16.4, which is observed twice by ultrasonography. They subsequently treated this value as unknown, with corresponding y_{0j} values of $y_{01} = 10.3$ and $y_{02} = 17.3$. Then, they estimate x_0 using their skewed Normal calibration model and compared this with the standard Normal calibration model. We did the same using the RS *GλD* calibration model and our results are shown in Table 5.

Table 4. Measurements obtained by dimensional measurement with a caliper (DM) and by ultrasonography (US) from the right testis for 42 teenagers, in ml^3

DM	US	DM	US	DM	US	DM	US	DM	US	DM	US
5.9	5	**17.3**	**16.4**	7.2	6.7	4.8	5.7	17.3	17.6	5.9	5.3
6.8	7.4	7.9	10	16.3	20	3.1	2.6	4.4	4.1	16.3	18.8
5	5.7	11.4	12.7	12.2	13.9	4.4	6.1	4.1	2.7	10.3	9.4
6	6.2	11.1	10.2	10.8	9.1	8.8	10.4	15.3	16.5	13	14.1
7.9	9.1	3.9	4.5	8.4	9.3	13	14.8	4.5	5.6	22.1	20.9
10.3	**16.4**	9.7	11	10.6	11.5	8.2	9.6	11.3	9.2	9.7	9.7
19.8	15.7	8.8	8.5	11.6	13.7	2	3	6.1	5.4	8.1	8.9

Table 5. A comparison of linear calibration models

	RS *GλD* model		*SN* model		Normal model	
Parameter	Estimate	Stdev.	Estimate	Stdev.	Estimate	Stdev.
α	0.014	0.497	-0.69	-	0.32	0.56
β	0.855	0.035	0.86	0.07	0.92	0.05
σ	-	-	2.13	-	1.55	0.17
x_0	12.128	0.963	12.66	1.81	14.58	1.24
λ	-	-	2.16	1.73	-	-
λ_2	0.146	0.355	-	-	-	-
λ_3	-0.030	0.061	-	-	-	-
λ_4	-0.162	0.184	-	-	-	-
AIC	150.36		160.69		163.74	
BIC	160.79		169.38		170.69	
HQ	144.58		156.55		161.15	

The theoretical derivation of the variability of our estimates under RS *GλD* is not readily tractable as in the cases of skewed Normal and Normal distributions. As we need to numerically derive our calculations, small errors in numerical procedures could accumulate into large errors even if we could evaluate the exact theoretical solution. As a workaround, we adopt the following procedure. Once we obtained the parameters of our model, α, β, x_0, λ_2, λ_3, λ_4, we conduct simulations to estimate the variability of our estimate. We use our estimated parameters from the RS $G\lambda D$ calibration model and x_i (excluding $x_i = 16.4$) from the original data to randomly generate y_{0j} and y_i according to (3.1) and (3.2). We then maximise the likelihood in (3.6) using Nelder Mead Simplex algorithm with initial values being our original estimated parameters. We repeat the process 1000 times and calculate the sample standard deviations of our estimated parameters.

Table 5 lists the estimated parameters and their standard deviations from RS $G\lambda D$, skewed Normal and Normal calibration models. We compute the Akaike, Bayesian and Hannan-Quinn information criterion (AIC, BIC, and HQ) to allow model selection between three models. All three criterion favors the RS $G\lambda D$ calibration model. In addition, the RS $G\lambda D$ model is much more efficient compared to the other models, with the smallest variability in its parameter estimates.

6. Concluding Remarks

We propose a new calibration model with RS $G\lambda D$ errors, which is an extremely flexible model that can cope with a wide range of different error distributions. Our method also lends to the development of FKML $G\lambda D$ calibration model, which may have better properties with regard to numerical convergence. Our simulations studies suggest our proposed model perform well for small sample sizes across a range of inverse scale and shape parameters of RS $G\lambda D$. We further demonstrate that the RS $G\lambda D$ calibration model can outperform skewed Normal or Normal calibration model, with lower AIC, BIC and HQ information criterion and lower variability in our parameter estimates in the context of a real life data. These simulation results are promising and future statistical models should aim to develop statistical technique that are tailored to data, rather than requiring empirical data to satisfy a particular statistical model. One possible extension of our model is the development of a mixture RS $G\lambda D$ calibration model, which would extend the flexibility of our model even further but also present a very challenging problem for data with small samples.

References

Azzalini, A. (1985). A class of distributions which includes the normal one. *Scandinavian Journal of Statistics, 12*, 171-178.

Chipkevitch, E., Nishimura, R., Tu, D., & Galea-Rajas, M. (1996). Clinical measurements of testicular volume in adolescents: Comparison of the reliability of 5 methods. *The Journal of Urology, 156*, 2050-2053. http://dx.doi.org/10.1016/S0022-5347(01)65433-8

Eisenhart, C. (1939). The interpretation of certain regression methods and their use in biological and industrial research. *Annals of Mathematical Statistics, 10*, 162-186. http://dx.doi.org/10.1214/aoms/1177732214

Figueiredoa, C., Bolfarinea, H., Sandovala, M., & Limab, C. (2010). On the skew-normal calibration model. *Journal of Applied Statistics, 37*(3), 435-451. http://dx.doi.org/10.1080/02664760802715906

Freimer, M., Kollia, G., Mudholkar, G. S., & Lin, C. T. (1988). A study of the generalised tukey lambda family. *Communications in Statistics-Theory and Methods, 17*, 3547-3567. http://dx.doi.org/10.1080/03610928808829820

Karian, Z. A., & Dudewicz, E. J. (2000). *Fitting statistical distributions: The generalized lambda distribution and generalised bootstrap methods*. New York: Chapman and Hall. http://dx.doi.org/10.1201/9781420038040

Ramberg, J. S., & Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the Association for Computing Machinery, 17*, 78-82. http://dx.doi.org/10.1145/360827.360840

Su, S. (2007a). Fitting single and mixture of generalised lambda distributions to data via discretized and maximum likelihood methods: GLDEX in R. *Journal of Statistical Software, 21*(9).

Su, S. (2007b). Numerical maximum log likelihood estimation for generalized lambda distributions. *Computational Statistics and Data Analysis, 51*(8), 3983-3998. http://dx.doi.org/10.1016/j.csda.2006.06.008

Su, S. (2010). Handbook of distribution fitting methods with R. In E. Karian, & Z. Dudewicz (Eds.), *Fitting GLD to data Using the GLDEX 1.0.4 in R* (Chap. 15). CRC Press.

ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Improved Measure on Extended Marginal Homogeneity for Ordinal Square Contingency Tables

Kouji Yamamoto[1], Ryota Shinjo[2] & Sadao Tomizawa[2]

[1] Department of Medical Innovation, Osaka University Hospital, Yamadaoka, Suita, Osaka, Japan

[2] Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Yamazaki, Noda City, Chiba, Japan

Correspondence: Kouji Yamamoto, Department of Medical Innovation, Osaka University Hospital, Yamadaoka, Suita, Osaka 565-0871, Japan. E-mail: yamamoto-k@hp-crc.med.osaka-u.ac.jp

Received: October 7, 2012 Accepted: January 15, 2013 Online Published: January 22, 2013

doi:10.5539/ijsp.v2n1p108 URL: http://dx.doi.org/10.5539/ijsp.v2n1p108

Abstract

For square contingency tables with ordered categories, Yamamoto et al. (2007) considered a measure to represent the degree of departure from extended marginal homogeneity. It attains the maximum value when one of two symmetric cumulative probabilities is zero. The present paper proposes an improved measure so that the degree of departure from extended marginal homogeneity can attain the maximum value even when the cumulative probabilities are not zeros. An example is given.

Keywords: marginal homogeneity, measure, Patil-Taillie diversity index, Shannon entropy

1. Introduction

For the $R \times R$ square contingency table, let π_{ij} denote the probability that an observation will fall in cell (i, j) $(i = 1, \dots, R;\ j = 1, \dots, R)$. The marginal homogeneity (MH) model is defined by

$$\pi_{i\cdot} = \pi_{\cdot i} \quad (i = 1, \dots, R),$$

where $\pi_{i\cdot} = \sum_{k=1}^{R} \pi_{ik}$ and $\pi_{\cdot i} = \sum_{k=1}^{R} \pi_{ki}$ (Stuart, 1955; Bishop et al., 1975, p. 294). Let

$$H_{1(i)} = \sum_{s=1}^{i} \sum_{t=i+1}^{R} \pi_{st}, \quad H_{2(i)} = \sum_{s=i+1}^{R} \sum_{t=1}^{i} \pi_{st},$$

for $i = 1, \dots, R-1$. This model may be expressed as

$$H_{1(i)} = H_{2(i)} \quad (i = 1, \dots, R-1).$$

This states that the cumulative probability that an observation will fall in row category i or below and column category $i+1$ or above is equal to the cumulative probability that the observation falls in column category i or below and row category $i+1$ or above for $i = 1, \dots, R-1$.

Tomizawa (1984, 1995) considered the extended marginal homogeneity (EMH) model which is expressed as

$$H_{1(i)} = \delta H_{2(i)} \quad (i = 1, \dots, R-1).$$

When $\delta = 1$, this is the MH model. Let

$$H_1 = \sum_{i=1}^{R-1} H_{1(i)}, \quad H_2 = \sum_{i=1}^{R-1} H_{2(i)}.$$

Assume that $\{H_{1(i)} + H_{2(i)} > 0\}$, $H_1 > 0$, and $H_2 > 0$. The EMH model may also be expressed as

$$Q_{1(i)} = Q_{2(i)} \quad (i = 1, \dots, R-1),$$

where

$$Q_{1(i)} = \frac{H^*_{1(i)}}{H^*_{1(i)} + H^*_{2(i)}}, \quad Q_{2(i)} = \frac{H^*_{2(i)}}{H^*_{1(i)} + H^*_{2(i)}},$$

$$H^*_{1(i)} = \frac{H_{1(i)}}{H_1}, \quad H^*_{2(i)} = \frac{H_{2(i)}}{H_2}.$$

This indicates that there is a structure of symmetry between $\{Q_{1(i)}, Q_{2(i)}\}$. Yamamoto et al. (2007) considered a measure to represent the degree of departure from EMH, using Patil and Taillie (1982) diversity index. The measure ranges between 0 and 1, and the degree of departure from EMH is maximum when $Q_{1(i)} = 0$ or $Q_{2(i)} = 0$ for all $i = 1, \ldots, R-1$. [Note that for measures for other models, e.g., the symmetry model (Bowker, 1948) and the MH model, see (e.g., Tomizawa et al., 2001; Tahata et al., 2006; Tahata et al., 2009)].

However, for analyzing square contingency tables, all $Q_{1(i)}$ and $Q_{2(i)}$ $(i = 1, \ldots, R-1)$ are positive in many cases. Thus, then Yamamoto et al. (2007) measure cannot attain the maximum value. So, we are now interested in a measure to represent the degree of departure from EMH such that it can attain the maximum value even when each of $\{Q_{1(i)}\}$ and $\{Q_{2(i)}\}$ is not zero.

For square contingency tables with ordered categories, the present paper proposes such a measure on EMH when all cumulative probabilities are positive.

2. New Measure

Let

$$E_i = \frac{H^*_{1(i)} + H^*_{2(i)}}{2} \quad (i = 1, \ldots, R-1).$$

For a specified d with $0.5 < d \leq 1$ and $1 - d \leq Q_{1(i)} \leq d$ $(i = 1, \ldots, R-1)$, define the new measure as, for $\lambda(> -1)$ fixed,

$$\Omega = \frac{1}{K}\left(1 - \frac{\lambda 2^\lambda}{2^\lambda - 1}\sum_{i=1}^{R-1} E_i W_i\right),$$

where

$$K = 1 - \frac{\lambda 2^\lambda}{2^\lambda - 1} L,$$

$$L = \frac{1}{\lambda}\left(1 - d^{\lambda+1} - (1-d)^{\lambda+1}\right),$$

$$W_i = \frac{1}{\lambda}\left(1 - Q_{1(i)}^{\lambda+1} - Q_{2(i)}^{\lambda+1}\right),$$

and the value at $\lambda = 0$ is taken to be continuous limit as $\lambda \to 0$. Thus, when $\lambda = 0$,

$$\Omega = \frac{1}{K}\left(1 - \frac{1}{\log 2}\sum_{i=1}^{R-1} E_i W_i\right),$$

where

$$K = 1 - \frac{1}{\log 2} L,$$

$$L = -d\log d - (1-d)\log(1-d),$$

$$W_i = -Q_{1(i)}\log Q_{1(i)} - Q_{2(i)}\log Q_{2(i)}.$$

Note that W_i is Patil-Taillie diversity index including Shannon entropy (when $\lambda = 0$). A value of d is chosen by the user such that $1 - d \leq Q_{1(i)} \leq d$ for any $i = 1, \ldots, R-1$. When $d = 1$, the measure Ω is identical to Yamamoto et al. (2007) measure. [Although the detail is omitted, note that Ω can also be expressed by using the power-divergence.]

Then, we can obtain the following theorem:

Theorem 1 *For each λ and a fixed d,*

(i) $0 \leq \Omega \leq 1$,

(ii) $\Omega = 0$ *if and only if the EMH model holds,*

(iii) $\Omega = 1$ *if and only if the degree of departure from EMH is the largest in the sense that* $Q_{1(i)} = d$ *or* $Q_{2(i)} = d$ *for all* $i = 1, \ldots, R-1$.

Proof. When $d = 1$, for each λ, the minimum value of W_i is 0 when $Q_{1(i)} = 0$ or $Q_{2(i)} = 0$ for all $i = 1, \ldots, R-1$, and the maximum value of it is $(2^\lambda - 1)/(\lambda 2^\lambda)$ (if $\lambda \neq 0$) or $\log 2$ (if $\lambda = 0$), when $Q_{1(i)} = Q_{2(i)} = 1/2$ for all $i = 1, \ldots, R-1$. When $d \neq 1$, the minimum value of it is L, which is not equal to 0, and the maximum value of it is the same as $d = 1$. Thus, the measure Ω lies between 0 and 1. So the proof is completed.

We note that the measure Ω is the modified measure of Yamamoto et al. (2007) by using a coefficient $1/K$.

Consider the artificial 4×4 table data in Table 1a on cell probabilities $\{p_{ij}\}$. Then, we see the degree of departure from EMH by using the existing measure Ω with $d = 1$ (i.e., Yamamoto et al. measure) and the measure Ω with $d < 1$ (in this case we set d= 0.9). We see from Table 1b that the true value of Ω with $d = 1$ is 0.531 (when $\lambda = 0$), and that of Ω with $d = 0.9$ is 1 (when $\lambda = 0$). Thus, we can see that the new measure Ω with $d < 1$ attains the maximum value 1, though all cumulative probabilities are positive.

Table 1. (a) An artificial 4×4 table data on cell probabilities $\{p_{ij}\}$, and (b) the values of measure Ω with $d = 1$ (existing measure) and Ω with $d = 0.9$ (new measure) applied to Table 1a

(a) Artificial data

	(1)	(2)	(3)	(4)
(1)	0.2	0.00025	0.00025	0.0005
(2)	0.003	0.2	0.089	0.00025
(3)	0.003	0.001	0.2	0.00825
(4)	0.003	0.003	0.075	0.2135

(b) Value of the existing measure and new measure

Existing measure	New measure
0.531	1

3. Asymptotic Variance for Estimated Measure

Let n_{ij} denote the observed frequency in cell (i, j) $(i = 1, \ldots, R; j = 1, \ldots, R)$. Assuming a multinomial distribution, the estimated measure $\hat{\Omega}$ is given by Ω with $\{\pi_{ij}\}$ replaced by $\{\hat{\pi}_{ij}\}$, where $\hat{\pi}_{ij} = n_{ij}/n$ and $n = \sum\sum n_{ij}$. Using the delta method, $\hat{\Omega}$ has asymptotically (as $n \to \infty$) a normal distribution with mean Ω and variance

$$\sigma^2 = \frac{1}{nK^2} \sum_{k=1}^{R-1} \sum_{l=k+1}^{R} \left[\pi_{kl} (v_{1(kl)})^2 + \pi_{lk} (v_{2(kl)})^2 \right],$$

where for $\lambda \neq 0$,

$$v_{s(kl)} = \frac{2^\lambda}{2(2^\lambda - 1) H_s} \left[\sum_{i=k}^{l-1} \tau_{s(i)} - (l-k) \sum_{i=1}^{R-1} H^*_{s(i)} \tau_{s(i)} \right] \quad (s = 1, 2),$$

with

$$\begin{aligned} \tau_{1(i)} &= (Q_{1(i)})^\lambda + \lambda \left\{ (Q_{1(i)})^\lambda - (Q_{2(i)})^\lambda \right\} Q_{2(i)}, \\ \tau_{2(i)} &= (Q_{2(i)})^\lambda + \lambda \left\{ (Q_{2(i)})^\lambda - (Q_{1(i)})^\lambda \right\} Q_{1(i)}, \end{aligned}$$

and for $\lambda = 0$,

$$v_{s(kl)} = \frac{1}{2H_s(\log 2)} \left[\sum_{i=k}^{l-1} \log Q_{s(i)} - (l-k) \sum_{i=1}^{R-1} H^*_{s(i)} \log Q_{s(i)} \right] \quad (s = 1, 2).$$

Let $\hat{\sigma}^2$ denote σ^2 with $\{\pi_{ij}\}$ replaced by $\{\hat{\pi}_{ij}\}$. Using these, the approximate confidence interval for the measure Ω is obtained as follows:

$$\hat{\Omega} \pm Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}},$$

where $Z_{\alpha/2}$ is the $(1 - \alpha/2)$ percentile of the standard normal distribution.

4. An Example

Consider the data in Table 2, taken from Hattori et al. (2002, p. 244). These data describe the cross-classification of father's and son's occupational status categories in Japan which were examined in 1955 and in 1975.

Table 2. Occupational status for Japanese father-son pairs (from Hattori et al., 2002, p. 244)

(a) Examined in 1955

	Son's status				
Father's status	(1)	(2)	(3)	(4)	Total
(1)	59	41	18	13	131
(2)	45	136	70	27	278
(3)	25	75	236	43	379
(4)	62	131	212	686	1091
Total	191	383	536	769	1879

(b) Examined in 1975

	Son's status				
Father's status	(1)	(2)	(3)	(4)	Total
(1)	127	101	54	12	294
(2)	86	207	125	13	431
(3)	78	124	310	24	536
(4)	109	206	437	325	1077
Total	400	638	926	374	2338

Note: (1) is Upper White-collar; (2) Lower White-collar; (3) Blue-collar and (4) Farming.

It seems natural to assume that all cumulative probabilities are positive because any observations can fall in all cells of the table. Therefore, it may not be appropriate to use the measure Ω with $d = 1$ because there is not a structure of cumulative probabilities such that Ω with $d = 1$ attains the maximum value 1. So we should use Ω with $d < 1$ (for example, $d = 0.99$) so that the measure can attain the maximum value 1.

Since the confidence intervals for Ω with $d = 0.99$ applied to the data in each of Tables 2a and 2b, do not include zero for all λ (see Table 3), these would indicate that there is not a structure of EMH in neither of tables.

Table 3. When $d = 0.99$, the estimate of Ω, estimated approximate standard error (S.E.) for $\hat{\Omega}$, and approximate 95% confidence interval (C.I.) for Ω, applied to Tables 2a and 2b

	λ	$\hat{\Omega}$	S.E.	C.I.
For Table 2a	−0.5	0.023	0.007	(0.010, 0.036)
	0.0	0.033	0.009	(0.014, 0.051)
	0.5	0.039	0.011	(0.018, 0.061)
	1.0	0.043	0.012	(0.019, 0.067)
	1.5	0.044	0.012	(0.020, 0.068)
	2.0	0.043	0.012	(0.019, 0.067)
	2.5	0.041	0.012	(0.018, 0.063)
For Table 2b	−0.5	0.105	0.012	(0.080, 0.129)
	0.0	0.141	0.016	(0.110, 0.172)
	0.5	0.165	0.017	(0.131, 0.199)
	1.0	0.177	0.018	(0.141, 0.213)
	1.5	0.180	0.018	(0.144, 0.216)
	2.0	0.177	0.018	(0.141, 0.213)
	2.5	0.170	0.018	(0.135, 0.205)

Moreover, we compare the degree of departure from EMH in Tables 2a and 2b using the confidence intervals for Ω. For any λ, the values in the confidence interval for Ω applied to the data in Table 2b are greater than those applied to the data in Table 2a. In addition, the values in the confidence interval do not overlap for Table 2a and for Table 2b. Thus, the degree of departure from EMH is greater for Table 2b than for Table 2a.

5. Concluding Remarks

We have proposed Ω which is an improvement of Yamamoto et al. (2007) measure (i.e., Ω with $d = 1$) to represent the degree of departure from EMH. For analyzing the data of square table such that all cumulative probabilities are positive, it may not be adequate to use the measure Ω with $d = 1$ because then the measure cannot attain the maximum value 1. For such data, it would be natural to use the measure Ω with $d < 1$ because then the measure can attain maximum value 1 even when all cumulative probabilities are positive.

The analyst may also be interested in how the value of d is determined. However it seems difficult to discuss this. The measure Ω depends on the value of a fixed d. Also, the value of Ω increases as the value of d decreases. But when we compare several tables, the result of comparisons is invariant without depending on the value of d. For analyzing a square table data, we note that if $1 - d \leq Q_{1(i)} \leq d$ is not satisfied for all $i = 1, \dots, R - 1$, the measure Ω cannot be used for the given data. Thus, the analyst must set the value of d carefully, so as to satisfy the condition $1 - d \leq Q_{1(i)} \leq d$ for all $i = 1, \dots, R - 1$. Therefore we recommend a value being close to 1 (for example, $d = 0.99$) as the value of d.

Acknowledgments

The authors would like to thank the editor and two anonymous reviewers for the helpful comments and suggestions.

References

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: The MIT Press.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association, 43*, 572-574. http://dx.doi.org/10.1080/01621459.1948.10483284

Hattori, T., Funatsu, T., & Torii, T. (2002). *Ajia Chukanso no Seisei to Tokushitsu (The Emergence and Features of the Asian Middle Classes).* The Institute of Developing Economies, Chiba, Japan (in Japanese).

Patil, G. P., & Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association ,77*, 548-561. http://dx.doi.org/10.1080/01621459.1982.10477845

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika, 42*, 412-416. http://dx.doi.org/10.1093/biomet/42.3-4.412

Tahata, K., Iwashita, T., & Tomizawa, S. (2006). Measure of departure from symmetry of cumulative marginal probabilities for square contingency tables with ordered categories. *SUT Journal of Mathematics, 42*, 7-29. Retrieved from http://www3.ma.kagu.tus.ac.jp/sutjmath/_userdata/42-1/02-tomizawa.pdf

Tahata, K., Yamamoto, K., Yamada, A., & Tomizawa, S. (2009). Generalized measures of departure from symmetry for square contingency tables. *Behaviormetrika, 36*, 75-86. http://dx.doi.org/10.2333/bhmk.36.75

Tomizawa, S. (1984). Three kinds of decompositions for the conditional symmetry model in a square contingency table. *Journal of the Japan Statistical Society, 14*, 35-42. Retrieved from http://www.jss.gr.jp/ja/journal/jjss1984.html

Tomizawa, S. (1995). A generalization of the marginal homogeneity model for square contingency tables with ordered categories. *Journal of Educational and Behavioral Statistics, 20*, 349-360. http://dx.doi.org/10.3102/10769986020004349

Tomizawa, S., Miyamoto, N., & Hatanaka, Y. (2001). Measure of asymmetry for square contingency tables having ordered categories. *Australian and New Zealand Journal of Statistics, 43*, 335-349. http://dx.doi.org/10.1111/1467-842X.00180

Yamamoto, K., Furuya, Y., & Tomizawa, S. (2007). Measure of departure from extended marginal homogeneity for square contingency tables with ordered categories. *REVSTAT: Statistical Journal, 5*, 269-283. Retrieved from http://www.ine.pt/revstat/pdf/rs070303.pdf

International Journal of Statistics and Probability; Vol. 2, No. 1; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Estimating Parameters from Samples: Shuttling between Spheres

Theodosia Prodromou[1]

[1] School of Education, University of New England, Australia

Correspondence: Theodosia Prodromou, School of Education, University of New England, Australia. Tel: 61-2-6773-3237. E-mail: theodosia.prodromou@une.edu.au

Received: December 31, 2012 Accepted: January 24, 2013 Online Published: January 27, 2013

doi:10.5539/ijsp.v2n1p113 URL: http://dx.doi.org/10.5539/ijsp.v2n1p113

Abstract

In order to better understand the thinking of students' learning to make informal statistical inferences, this research examined the thinking of senior secondary school students (age 17) engaged in the task of using observed data to make point estimates of a population parameter within a computer-based simulation. Following the "Growing Samples" instructional model, the point estimation activity involved sampling and estimating across three tasks with different sample sizes. This research study aimed to trace the evolution of the students' thinking, with particular attention to use of the statistical concepts in making informal inferences from sampling. The students in this study were observed to rely primarily on mathematical thinking, which, perhaps, inhibited their ability to construct meanings about the basic statistical concepts underpinning sampling when performing point estimates. At times in the process students were seen to shift between mathematical thinking, statistical thinking, and thinking about the context, but the mathematical thinking seemed to dominate their attempts to create estimates. These research findings are useful for informing the teaching of point estimation of a population parameter to school-aged students. The research findings also stress the need for teachers to rethink the relationship between statistical thinking and mathematical thinking in order to promote statistical thinking in relevant learning situations for their students.

Keywords: informal statistical inference, point estimates, population, samples, statistical thinking, mathematical thinking

1. Introduction

A productive and authentic way of teaching the statistical reasoning necessary when working with samples is to provide opportunities for school students to engage in activities that involve informal inferential reasoning (Makar, Wells, & Allmond, 2011). Such activities also provide an important way for students to progress from working with descriptive statistics to working with inferential statistics because they offer the opportunity to reason informally. The word "informally" is used to "emphasize that we are not expecting students to rely on formal statistical measures and procedures to formulate their inferences" (Makar, Bakker, & Ben-Zvi, 2011, p. 153). Such Informal Inferential Reasoning (IIR) has been defined as the process of drawing generalised conclusions from data, involving four critical principles: generalising beyond data (parameter estimates, conclusions, and predictions); using data as evidence of the generalisation; articulating the degree of certainty (due to variability) embedded in the generalisation (these three principles were articulated by Makar and Rubin, 2009); and comparing datasets with a model such as ideal (targeted) distributions (proposed by Bakker, Kent, Derry, Noss, and Hoyles, 2008).

The research reported in this paper focuses on the first of these four principles, generalising beyond data, in particular estimating parameters. Estimating parameters is a process by which one makes inferences about a population based on information gained from one (or more) sample(s). A sample is a representative part of a population selected when sampling for the purpose of drawing inferences about unknown populations (e.g., estimating parameters or predicting).

2. Growing Samples

Recent research has sought to understand how better to approach the topic of making informal inferences about a population based on information gained from one or more samples (Ben-Zvi et al., 2011, 2012; Prodromou, 2011).

The instructional idea of "Growing Samples", suggested by Konold and Pollatsek (2002) and then developed by Bakker (2004), plays a predominant role in providing a useful perspective of the role in shedding some light in the development of students' informal inferential reasoning, and reasoning about sampling, samples, and variation.

Bakker helped eighth grade students who engaged with a sequence of "Growing Samples" activities to see stable patterns generated by larger samples, thus students understood that larger samples are less variable and better represent population. Bakker suggested that asking students to make conjectures about the growing samples builds students' reasoning about sampling in the context of variability and distribution. Research literature provides evidence that the growing samples approach is helpful in supporting coherent reasoning, based exclusively on the integration of key statistical concepts such as sampling, data, distribution, variability, and tendency (Ben-Zvi et al., 2011, 2012; Prodromou, 2011).

Ben-Zvi (2006) found that the growing samples processes enhanced students' sensitivity to uncertainty and variation in data, enabling students to know something about the population. Research studies by Ben-Zvi et al. (2011) and Prodromou (2011, 2012), which were in line with the Growing Samples literature, showed that students developed inferential reasoning about sampling while working with *TinkerPlots*. The students in those studies not only experienced the limitations of small samples when making inferences about a larger population, but also experienced an emerging quantification of confidence in making such inferences, interconnections of concepts of sampling, and informal statistical inference with key concepts such as spread, distribution, likelihood, randomness, average, and graph interpretation.

When the students were encouraged to express their confidence about how certain they were about their inferences, they tended to either express extreme confidence in knowing that something can be inferred from samples or express that nothing could be concluded (i.e., complete certainty vs. extreme doubt; Ben-Zvi et al., 2012). The growing samples task design provided students with opportunities to witness increasing evidence for (or against) particular conjectures, thus develop a language to talk about "grey areas of this middle ground" (p. 923).The research reported in this paper is based on the growing samples design of the investigations that support students' informal inferential reasoning when estimating population parameters from samples.

3. Using Samples to Estimate Parameters as Part of Informal Inferential Reasoning

Parameters can be estimated by providing either a point estimate or an interval estimate. A point estimate involves the use of sample data to calculate a single value (best known as a statistic) that can be used as a "best guess" or "best estimate" of an unknown (fixed or random) population parameter. For example, a sample mean is a point estimate used to estimate the population mean. An "interval estimate" involves the use of sample data to calculate an interval of possible (or probable) values of an unknown population parameter within which a population parameter lies. For example, 1< sample mean < 4 is an interval estimate within which the population mean lies.

The student reasoning process that leads to informal statistical inferences (ISI) when estimating parameters can help teachers to gain insights in student thinking and identify critical elements that support and nurture student ISI such as estimating parameters. Statistical thinking is needed when students engage in informal inferential reasoning and teachers need to be familiar with this type of thinking and to nurture it for students to be supported in their informal inferential reasoning.

4. Tension between Statistical Thinking and Mathematical Thinking

Mathematics teachers need to be aware that the thinking a student requires to solve a statistical problem will differ from the thinking required to solve most mathematical problems. If students are not equipped with sufficient statistical thinking capabilities they may approach statistical problems using mathematical thinking. So how might statistical thinking and mathematical thinking differ?

What is 'thinking'? Generally speaking, thinking can be defined as "the process of considering or reasoning about something" (Oxford University Press, 2012). While this definition provides a basic guide to the concept of thinking, more detailed explanations have been provided that are discipline specific, of these the statistical and mathematical thinking are of most interest. Before considering these in more detail, what of the distinction between thinking and reasoning?

In statistics education some have described reasoning as a form of thinking with both reasoning and other forms of thinking needed to be able to work on a task, while others have attempted to make a clear distinction reasoning and thinking. A useful approach to distinguishing between reasoning and thinking is to consider the task being undertaken and conceptualise thinking as *knowing* "when and how to apply knowledge and procedures", and reasoning as *explaining* "why results were produced or why a conclusion was justified" (delMas, 2004, p. 85).

Thus examples of reasoning can be found in particular stages of a person's thinking, such as where the person is expected to imply, justify, or infer. Now back to the two types of thinking, mathematical and statistical.

What is mathematical thinking? Mason, Burton, and Stacey (2010) described four fundamental processes involved in mathematical thinking: (MT1) specialising-considering special cases or examples; (MT2) generalising-looking for patterns and relationships; (MT3) conjecturing-predicting relationships and results; and (MT4) convincing-finding and communicating reasons why something is true. From the previous discussion it might be concluded that convincing (MT4) is mathematical thinking that involves "reasoning".

What is statistical thinking? In attempting to answer this question, a statistician and a mathematics educator (Wild & Pfannkuch, 1999) worked together to build up four dimensions which contribute to the "rich complexity" of statistical thinking: (ST1) the investigative cycle-continuously through the stages problem, plan, data, analysis and conclusion; (ST2) types of thinking-recognition of need for data, transnumeration, consideration of variation, reasoning with distinctive set of statistical models, integrating the statistical and contextual information, knowledge, and conceptions; (ST3) the interrogative cycle-continuously through the stages generate, seek, interpret, criticise and judge; and (ST4) dispositions-including scepticism, imagination, curiosity and awareness, openness to ideas that challenge preconceptions, a propensity to seek deeper meaning, being logical, engagement and perseverance.

Amongst the types of thinking skills (ST2), Wild and Pfannkuch recognised, in particular, the importance of the raw materials on which statistical thinking works. These raw materials are statistical knowledge, context knowledge, and the information in data. However, the thinking itself occurs by the synthesis of these elements. In particular, one has to bring to bear all appropriate knowledge regarding the undertaken task, and then to build connections amongst existing context-knowledge and the outcomes of statistical analyses. Wild and Pfannkuch (1999) described the synthesis of context-knowledge and statistical knowledge as one that "traces the (usual) evolution of an idea from the earliest inkling through to the formulation of a statistical question precise enough to be answered by the collection of data, and then on to a plan of action" (p. 228). They also emphasize the continual shuttling backwards and forwards between thinking in the context sphere and the statistical sphere. The interplay between context and statistics is continuous until the questions in hand are satisfactorily answered. For example, Wild and Pfannkuch (1999) explain how, in the analysis stage, context knowledge leads to questions that require consultation of the observation data, which pushes learners into the statistical sphere of thinking, but then characteristics of the data push learners back to the context sphere to answer basic questions like, "Why is this happening?", and "What does this mean?" (p. 228).

While there may be similarities between the mathematical thinking and statistical thinking, these two types of thinking are dissimilar in two important elements: variation and context. All statistical thinking must be grounded within a context (delMas, 2004), while mathematical thinking may or may not make use of contexts. All statistical thinking involves some form of consideration of variation (Pfannkuch & Wild, 2004), which is very different from the concept of variables dealt with in mathematical thinking. The fact that variation is an observable phenomena and that it is always present (Wild & Pfannkuch, 1999) is of relevance to all aspects of statistical thinking.

To appreciate the tension between thinking mathematically or statistically, consider research reported by Lane-Getaz (2006) where students engaged in simulation activities were good at "mathematically" calculating statistics but once they were exposed to activities that allowed them to explore variation within distributions they were able to produce explanations of their projects that demonstrated better statistical thinking.

Rather than promoting the differences between the two types of thinking, teachers who rethink the relationship between statistical thinking and mathematical thinking can help their students to learn how to synthesize the two types of thinking. This would help teachers to promote statistical thinking in relevant learning situations for their students. To inform teachers in developing such support, the researcher became interested in what type of thinking students will engage in when completing an activity that requires statistical thinking.

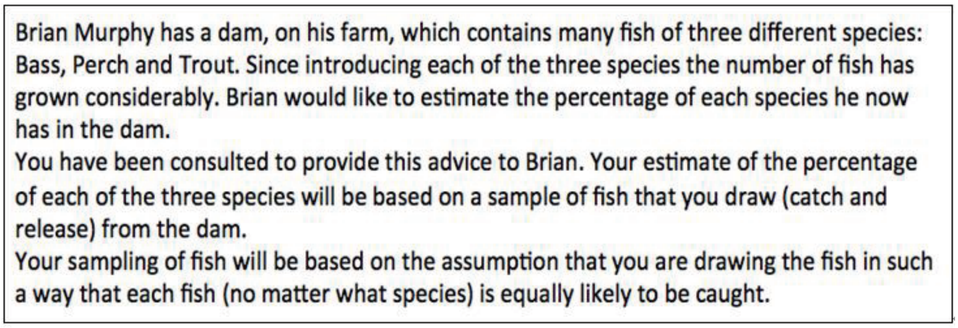
5. Aim

This exploratory research study examined how senior secondary school students construct meanings about basic statistical concepts underpinning sampling when making informal inferences from data. The focus was on observing the development of students' thinking as they construct meaning about the key statistical concepts of 'sample' and 'sampling,' while the students engaged in an informal statistical inference task that involved making point estimates of a population parameter within a computer-based simulation. It was expected that some insights might be gained into the conceptual struggle that takes place when 17-year-olds engage in inferential reasoning when making point estimates of a population parameter.

In this research a constructivist stance is used to search for nave conceptions that might serve as resources in developing more sophisticated strategies. In addition, this might shed some light on the tension that a student may experience when opting for thinking mathematically or statistically and how this tension can be resolved.

6. Point Estimation Activity

The point estimation activity, a computer simulation titled Murphy's Dam was presented in a spreadsheet and introduced the context of a dam containing three fish species (Bass, Perch, Trout; Figure 1). The spreadsheet allowed students to simulate drawing a sample of fish from the dam and displayed the number and percentage of each species of fish within the sample (Figure 2). The students were asked to provide the owner of the dam, Brian Murphy, with advice about the percentage of each species of fish in his dam. To inform their advice, the students were engaged in drawing "catch and release" samples of 20, 50, and then 100 fish from the dam (Tasks 1, 2, and 3, respectively). In each task students were requested to make estimates of the percentage of bass, perch and trout in the dam after each sample was drawn.

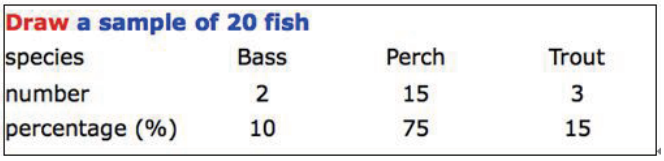


Brian Murphy has a dam, on his farm, which contains many fish of three different species: Bass, Perch and Trout. Since introducing each of the three species the number of fish has grown considerably. Brian would like to estimate the percentage of each species he now has in the dam.

You have been consulted to provide this advice to Brian. Your estimate of the percentage of each of the three species will be based on a sample of fish that you draw (catch and release) from the dam.

Your sampling of fish will be based on the assumption that you are drawing the fish in such a way that each fish (no matter what species) is equally likely to be caught.

Figure 1. Murphy's dam scenario

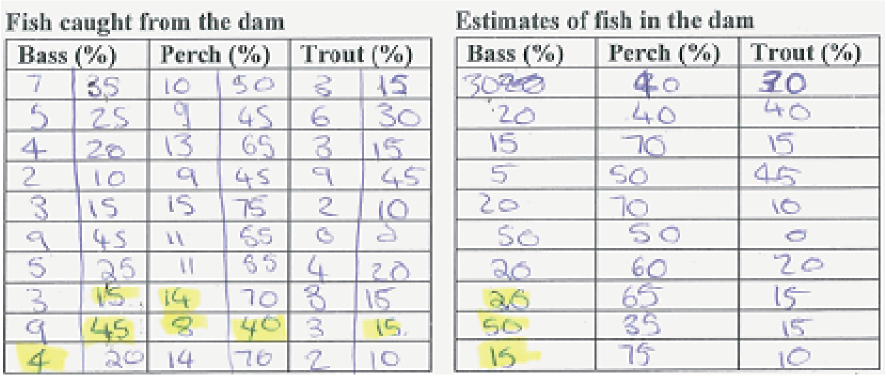


Draw a sample of 20 fish

species	Bass	Perch	Trout
number	2	15	3
percentage (%)	10	75	15

Figure 2. Simulated sample of 20 fish

For each task students were required to draw ten separate samples from the dam and for each sample drawn record the observed percentage for each species of fish caught and a point estimate of percentage of bass, perch and trout (Figure 3).



Fish caught from the dam

Bass (%)		Perch (%)		Trout (%)	
7	35	10	50	3	15
5	25	9	45	6	30
4	20	13	65	3	15
2	10	9	45	9	45
3	15	15	75	2	10
9	45	11	55	0	0
5	25	11	55	4	20
3	15	14	70	3	15
9	45	8	40	3	15
4	20	14	70	2	10

Estimates of fish in the dam

Bass (%)	Perch (%)	Trout (%)
30	40	30
20	40	40
15	70	15
5	50	45
20	70	10
50	50	0
20	60	20
20	65	15
50	35	15
15	75	10

Figure 3. Example of completed recording sheet (second iteration of Task 1)

When the three tasks were completed, the students were asked to reflect on the point estimation activities across the three tasks, reason about their estimates by comparing the estimates, and attempt to estimate the actual percentage of each species of fish in the dam (In the simulation, the percentage of fish in the dam was set as 30% bass, 50% perch and 20% trout).

The design of the point estimation activity evolved around the idea of growing samples, starting from a sample of size 20, moving to about 50, then 100, and finally the entire population. Using a sequence of "growing sample" activities was a pedagogical design conjecture to help students progressively develop their inferential reasoning about samples, and their ability to make point estimates of parameters of the population.

7. Methodology

7.1 Participants

The point estimation activity was undertaken by three pairs of average-ability female students studying Mathematics General (Year 11-age 17 years) in an Australian secondary school. Participation was voluntary; students self-selected a partner; and the tasks were undertaken out of class time. The teacher made the final choice of which students participated by recommending those who were able to better articulate their thinking. The choice of senior secondary students was a curriculum-based decision because, in accordance with Australian curriculum guidelines (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2011), Year 11 students in this school had been taught sampling prior to involvement in the research. The choice of average ability students was based on the assumption that above-average ability students would construct the knowledge too quickly and possibly not take the time to verbalise their thinking, while below-average ability students may not be able to articulate their thinking.

The researcher was a participant observer, working with each pair of students as they completed the three tasks. The researcher interacted with the students to probe for reasons that might help to explain their actions and therefore provide some insight in their thinking.

7.2 Data Analysis

Student work with the simulation was recorded using Camtasia (2000) software. The data collected included audio recordings of the student voices, video recordings of the screen activity, and worksheets completed by the students.

All the audio recordings were transcribed and screenshots of the simulation spreadsheet and student recording sheets were included as needed to make sense of the transcription. The researchers discussed the data and chose those sections of the transcript that most clearly demonstrated student thinking as they reasoned.

8. Findings

The findings are only reported for one pair of girls, Cathy and Liz (pseudonyms), because their articulation during the three tasks gives the most informative illustration of their thinking. When the two girls first engaged with the three tasks they simplified the required approach by estimating after all ten samples had been drawn rather than after each sample. A second iteration of the three tasks was undertaken to achieve what was originally planned for the point estimation activity, having the girls estimate after each sample was drawn. The findings are reported for each of the two iterations separately.

8.1 First Iteration of the Three Tasks

When Cathy and Liz were engaged in Task 1 (drawing samples of size 20), they began working on separate recording sheets (as instructed) but insisted on drawing all ten samples (contrary to instructions) from the dam before estimating the percentage for each species of fish. They made two requests: (i) to share one recording sheet rather than work on two separate sheets; and (ii) to record the number of fish as well as the percentage of fish for each species. Both requests were allowed. When making their estimate (after the ten samples were drawn) they insisted on trying to develop an algorithm to calculate the estimate. They took the average number (over the 10 samples) of bass caught (5) and divided this by the total number of fish caught (20) and converted it to a percentage (25%). They repeated this process to calculate the percentage of Perch (50%) and the percentage of Trout (25%). The working was done using a calculator. When they were asked to explain the algorithm they developed to make the estimate they wrote "number of fish caught sample of fish x 100" (Algorithm A). Using an algorithm like this is an example of mathematical thinking. They did not explain why they averaged the number of fish caught over all ten samples and then calculated the percentage, rather than just averaging the percentages.

When Cathy and Liz engaged in Task 2 (drawing samples of size 50), they tried to apply algorithm A to calculate the estimate of the percentage of bass. They drew the ten samples and performed their calculations, this time using a spreadsheet. The average of the ten samples came out to 14.8 but they chose to use 14 instead of 14.8 in their algorithm. In addition, they said they did not "like" the answer they were getting to estimate the percentage of bass because it was too small and they chose 20 instead. It is not clear how or why the girls chose 20% as their estimate. No percentages were calculated for perch and trout and the girls did not explain why they did not estimate the percentages of perch and trout.

When Cathy and Liz engaged in Task 3 (drawing a sample of size 100), they realised that in the observed data for each sample the percentage of each species of fish was equal to the number of fish caught. They again applied their algorithm to calculate the estimates. They wrote out the estimate for bass, and then argued that the estimates for the other two species of fish would simply equal the percentages they had drawn in their samples. As explained above the girls developed an algorithm to form their estimates and their explanations were computational rather than statistical in nature. The researchers decided to do a second iteration of the three tasks, this time insisting that the girls estimate the percentage of each species of fish after each sample drawn as was planned in the original task. The second iteration of the three tasks was performed two weeks after the first iteration.

8.2 Second Iteration of the Three Tasks

When Cathy and Liz engaged in Task 1 they experimented with a mathematical algorithm: "number of a species of fish divided by percentage of a species of fish multiplied by 100" (Algorithm B). For example, after drawing the first sample, they calculated 5 divided by 25 multiplied by 100 to give an estimate of 20 for the percentage of bass. Similarly for perch, 9 divided by 45 multiplied by 100 gave 20, and even for trout they calculated 20 as the estimate. Although they did not express surprise that all three estimates were the same, they did realise that the three percentages should sum to 100 (conservation principle) and as the sum was only 60, they concluded that that their algorithm did not work.

They then calculated the estimated percentage of each species using an alternate algorithm: "percentage of a species of fish divided by the number of species of fish multiply by 100" (Algorithm C). This calculation resulted in 50% as the estimate for each of bass, perch, and trout. This time they concluded that their algorithm did not work because they calculated the same estimate, 50, for each species of fish.

Not satisfied with either attempt at estimation, they suggested drawing another sample of 20 fish so that they could compare the new sample and the previous sample to observe any possible change. Cathy and Liz used the new and previous observed percentage caught to produce a new estimate, which was not related to the previous estimate. If the percentage of fish caught had gone up (down), then the estimated percentage of fish was set at 5% more (less) than the caught amount (Algorithm D). This algorithm was used for both bass and perch. Then the estimated percentage of trout was always calculated by adding the bass and perch estimates together and subtracting from 100. Algorithm D was the first attempt by the girls to provide estimates that in some way were linked to the fluctuations (variations) between the samples drawn but made no use of the previous estimate to calculate a new estimate. Obviously this algorithm only worked for the second or subsequent estimates, otherwise the change direct, up or down, could not be determined. The increments were always only 5%, up or down, irrespective of the size of the percentage of fish for the new and previous sample.

An example of the application of Algorithm D from the 8th estimates and 9th estimates of Task 2 (Figure 4) follows. To find the 9th estimate for bass Cathy and Liz noticed that the 9th catch (observed percentage) for bass, 45, was larger than the 8th, 15, and so the 9th estimate for bass was the observed value (45) plus 5, giving 50 as the estimated percentage of bass. Similarly for the 9th estimate for perch, 40 (new observed percentage) was less than 70 (previous observed percentage) and so 5 was subtracted from the observed percentage (40) to give 35 as the estimated percentage of perch. Finally, the 9th estimated percentage of trout was 15 was calculated as 100-(50+35). It should be noted that the 8th estimate (20 65 15) was not used at all by the students in producing the 9th estimate Although this application of Algorithm D has been explained using 8th and 9th catches, this algorithm was applied for calculating the second and subsequent estimates.

When Cathy and Liz were engaged in drawing "catch and release" samples of size 50 (Task 2), from the dam and estimating the percentage of bass, perch, and trout in the dam (see student recording sheet in Figure 5), they needed to form an estimate for the first sample drawn because Algorithm D could not be applied. Cathy tried Algorithm B, which resulted in 50 for each of the three percentage estimates, bass, perch, and trout. The girls realised that the estimate cannot be 50 every time, concluding that the algorithm did not work.

They tried to find a number that went into 28, 50, and 22 (the percentages caught in the first sample drawn). They concluded that 2 goes into 22, 50 and 28 and thus 2 was the number they could increase or decrease the observed percentage by the work out the percentage estimate.

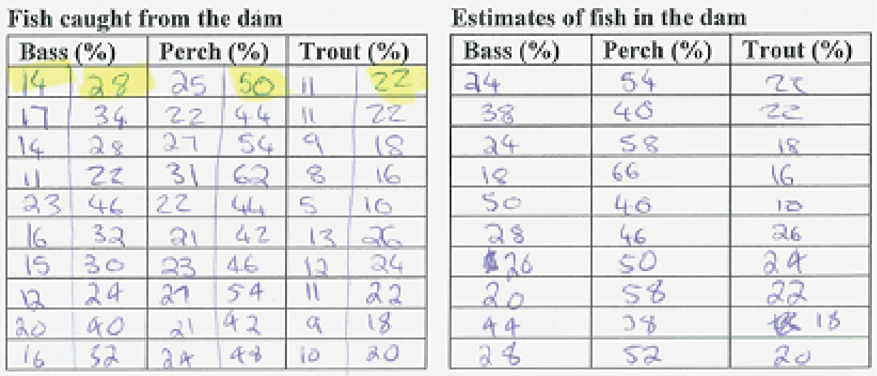
Fish caught from the dam

Bass (%)	Perch (%)	Trout (%)
15	70	15
45	40	15

Estimates of fish in the dam

Bass (%)	Perch (%)	Trout (%)
20	65	15
50	35	15

Figure 4. Example of the application of Algorithm D to estimate the percentages of each species of fish



Fish caught from the dam

Bass (%)		Perch (%)		Trout (%)	
14	28	25	50	11	22
17	34	22	44	11	22
14	28	27	54	9	18
11	22	31	62	8	16
23	46	22	44	5	10
16	32	21	42	13	26
15	30	23	46	12	24
12	24	27	54	11	22
20	40	21	42	9	18
16	52	24	48	10	20

Estimates of fish in the dam

Bass (%)	Perch (%)	Trout (%)
24	54	22
38	40	22
24	58	18
18	66	16
50	40	10
28	46	26
26	50	24
20	58	22
44	38	18
28	52	20

Figure 5. Student recording sheet for Task 2 in the second iteration. Each row represents a trial

The following comes from a point where they tried to make an estimate after the first sample was drawn.

1) Cathy: Maybe if we would work out like the next fish (draw another sample) because then we could see if there is a pattern which might help with the formula.

2) Researcher: Remember that you need to make estimates after each sample was drawn.

3) Liz: maybe if we could work out how many fish equalled what percentages like we did for Task 1 (sample size 20). 1 fish equalled 5% and we will work things out.

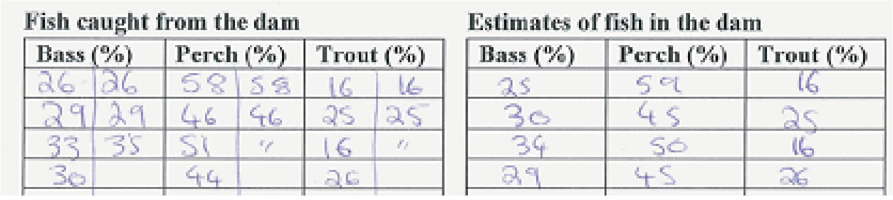
4) Cathy: Well it's always half. 1 fish equals 2%, 2 fish equals 4%, 4 fish equals 8%.

They had decided confidently that 2 was the number they could increase or decrease the observed percentage by the work out the percentage estimate.

The girls used then a new algorithm (Algorithm E), without discussing this algorithm. They went down by 2 fish for the Bass and so took away 4% from the "caught" percentage (28%) of Bass to give the "estimate" percentage (24%) of Bass. They then added 4% to the percentage of Perch. Then they found the percentage of trout by adding the percentage of bass and perch and subtracting the sum from 100%. They estimated [24, 54, 22] (See student recording sheet in Figure 5).

The girls tried to use Algorithm D developed for Task 1 (sample size 20) to make the second and subsequent estimates. However, when applying the algorithm this time the relevant change, either up or down, was only by 2% and not 5%. As before, the estimated percentage of trout was calculated using the conservation principle.

When Cathy and Liz were engaged in drawing "catch and release" samples of size 100 (Task 3), they applied Algorithm E, for the first estimate (see student recording sheet in Figure 6). However, they went down by 1 fish for the Bass and so took away 1% from the "caught" percentage (26%) of Bass to give the "estimate" percentage (25%) of Bass. They then added 1% to the percentage (58%) of Perch to give the estimated percentage (59%). Then they found the percentage of trout by adding the percentage of bass and perch and subtracting the sum from 100%.



Fish caught from the dam

Bass (%)		Perch (%)		Trout (%)	
26	26	58	58	16	16
29	29	46	46	25	25
33	35	51	"	16	"
30		44		26	

Estimates of fish in the dam

Bass (%)	Perch (%)	Trout (%)
25	59	16
30	45	25
34	50	16
29	45	26

Figure 6. Student recording sheet for Task 3 in the second iteration

They drew three samples and made three estimates based on Algorithm D, to make the second and subsequent estimates. They relevant change was either up or down, by 1%. Cathy pointed out that the algorithm was working "even" when the number of fish caught was 100 because the number of fish equals the percentage and so 1 fish equalled 1%. As before, the estimated percentage of trout was calculated using the conservation principle. They then concluded that they did not need to draw any more samples because it was "easy" to estimate.

When the girls were asked to reflect on all the activities when they "caught and released" samples of sizes 20, 50 and 100 and attempt estimate the actual percentage of each species of fish in the dam. They averaged the percentage of the fish caught when 100 fish caught from the dam. The researcher asked them to estimate the percentage without using a formula. Then they looked at the range of the percentage caught for a particular species across the three activities (see student recording sheets in Figure 3, Figure 5, and Figure 6), e.g., for bass 26 was the least and 33 was the most and then they found the median "middle value" (33-26=7) and then halved it so that the middle value was 3.5. They concluded that it could be 29 or 30 but then stayed with 30.

9. Summary

The conclusions drawn indicate when making point estimates of the percentage of bass, perch, and trout in the dam, the two secondary students, Cathy and Liz, generally focused on whether the percentage of fish caught had gone up (or down), then the estimated percentage of fish was set at 5% (samples of size 20), or 2% (samples of size 50), or 1% (samples of size 100) more (less) than the caught amount.

A number of algorithms emerged and as part of this process the students tended to experiment with choosing a "relevant" algorithm each time a new sample was drawn.

Despite the apparently arbitrary choice of the three following algorithms to forms estimates of the percentage of bass, perch and trout in the dam:

(1) "number of fish caught divided by sample of fish multiplied by 100" (Algorithm A).

(2) "number of a species of fish divided by percentage of a species of fish multiplied by 100" (Algorithm B)

(3) "percentage of a species of fish divided by the number of species of fish multiply by 100" (Algorithm C).

When Cathy and Liz engaged in Task 2 (drawing sample of size 50) and Task 3 (drawing sample of size 100), they tried to apply algorithm E to make the first estimate:

(4) "go down by 2 fish when drawing a sample of size 50 (or 1 fish drawing a sample of size 100) for the Bass and so take away $2 \cdot 2\% = 4\%$ (or 1%) from the "caught" percentage of Bass to give the "estimate" percentage of Bass. They then added 4% to the percentage of Perch. Then they found the percentage of trout by adding the percentage of bass and perch and subtracting the sum from 100%" (Algorithm E).

The girls tried to use Algorithm D to make the second and subsequent estimates:

(5) "the relevant change, either up or down, was by 5% (sample of size 20) or 2% (sample of size 50) or 1% (sample of size 100). The estimated percentage of trout was calculated using the conservation principle" (Algorithm D).

Eventually, students appeared to settle on Algorithm E when making the first point estimates of a population parameter and on algorithm D when making the second and subsequent estimates.

Whilst making the point estimates, the students demonstrated very little evidence of the notion of stabilizing the values used for the points estimate as more information becomes available with each successive sample. So, despite looking back to previous estimates to form the new estimate, there was no sense of refining previous estimates, just a sense of using them as a basis for the new estimate.

In this paper, students' engagement in the point estimation activity has been presented in which mathematical thinking dominates over statistical thinking.

Students did not construct any meanings about basic statistical concepts underpinning sampling in the context of making informal inferences from data when performing point estimates.

10. Discussion

For the contribution to the topic of estimating parameters from samples when engaged in an informal statistical inference task that involved making point estimates of a population parameter within a computer-based simulation, the research findings are useful for informing the teaching of point estimation of a population parameter to school-aged students. The type of thinking students engaged in when completing the point estimation activity that requires statistical thinking became the focus of the research findings. One has to bring to bear all relevant knowledge on the tasks in hand, and then to draw connections between existing context-knowledge, mathematical knowledge and the previous estimates of parameters from samples. Wild and Pfannkuch (1999) developed a theoretical framework that illustrates the Interplay between context and statistics and also emphasizes the continual shuttling backwards and forwards between thinking in the context sphere and the statistical sphere.

In this study, the findings show that the students persistence to think mathematically rather than statistically prevent them from constructing any meanings about basic statistical concepts underpinning sampling in the context of making informal inferences from data when performing point estimates. The research findings stress the need to rethink the relationship between statistical thinking and mathematical thinking.

The researcher considers that statistical knowledge, mathematical knowledge, and context knowledge are the raw materials on which thinking works. The thinking required for estimating the percentage of the population parameters is in fact a synthesis of these elements to produce implications, insights and conjectures. One cannot indulge in statistical thinking without having some context knowledge and mathematical knowledge.

The researcher based on Wild's and Pfannkuch' (1999) constructed a new framework that illustrates the construction of students' knowledge when engaged in an informal statistical inference task that involved making point estimates of population parameters from samples within a computer-based simulation. Figure 7 illustrates the continual shuttling backwards and forwards between thinking in the context sphere, the statistical sphere, and the mathematical sphere.

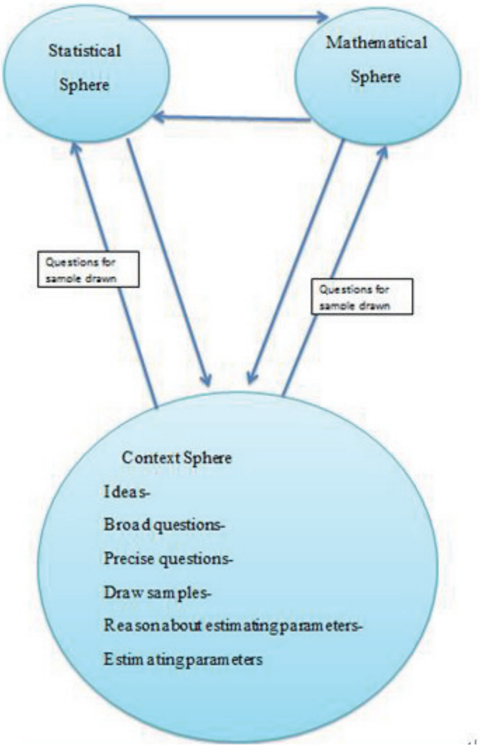


Figure 7. Shuttling between spheres

Figure 7 traces the evolution of an idea from the earliest inkling through to the formulation of a precise statistical question that is to be answered following the different stages of the “growing samples” instructional idea. The role of the context is crucial at the earliest stages of the statistical task because learners’ constructions of informal statistical inference understandings are driven almost entirely by context knowledge.

For example, at the parameter estimating stage questions are suggested by context knowledge that require consulting the samples drawn, which temporarily pushes learners into the mathematical sphere because learners are more familiar with constructing and working using mathematical formulae. Features seen in samples propel learner back to the context sphere to answer scaffolding questions: “Why is this happening?” and ”What does this mean?” Such scaffolding questions are meant to push learners into the statistical sphere to help them construct statistical concepts underpinning their activities. Statistical knowledge contributes more as the thinking crystallises.

In this study, students were not able to construct meanings about basic statistical concepts underpinning sampling in the context of making informal inferences from data when performing point estimates because the scaffolding questions did not push leaners into the statistical sphere.

It might be argued that students’ pursuit of mathematical thinking at the earliest stages of learning statistics has been the root of all the problems that have been encountered by students being taught elementary statistical concepts. At these earliest stages, the reduction of emphasis on mathematical thinking in statistical teaching is crucial because the earliest stages are driven almost entirely by context knowledge and learners’ thinking is moving backwards and forwards between thinking in the context sphere and the statistical sphere. The statistical concepts learners construct and reason about are usually “informal” and embrace elements of “naivety”.

A fundamental change in teaching elementary statistics at school is required in order to allow and facilitate the simplification of core statistical ideas. In particular, the key ideas of statistical inference can be developed without relying on any mathematical models used by formal probability theory. Statistical thinking at elementary levels can be developed relying on processes and not mathematics.

On the contrary, for constructing statistical knowledge beyond the elementary level, the use of mathematical formulae of probability theory becomes increasingly important, as mathematics is critical for the development of statistical methods.

The reader must consider the limitations of this research to elaborate the research questions. One limitation is that the researcher only analysed the reasoning of one pair of students, and focused on the more interesting illustrations of the emerging ideas. While the overall analysis of the other pairs of students followed the same broad strokes, some interesting variations used by the students in their reasoning might justify further discussion, but are outside the scope of this paper.

A second limitation is the interview technique used to ask students to explain their responses was not explicit enough. Despite students being asked “why” they had formed the estimates the way they did and “why” they made any changes to their estimates, they mostly gave superficial responses in their reasoning. More probing questions were needed to direct the students to explain their reasoning and propel students back to the context sphere to reflect on context knowledge.

10.1 Future Research

Despite the limitations of the study, it reveals some aspects of students’ reasoning while making point estimates. Although some interesting point estimation algorithms emerged, little evidence of using the core concepts, sampling, sample, and variation.

The results of this research provide a strong basis to help those teaching point estimation to secondary school students better understand their students’ reasoning. As far as further research is concerned, in raising the notion that there may be a better way to investigate statistical reasoning, especially as involved in the sampling process, it is acknowledged that the focus should not be on mathematical approaches to estimation.

There is still more research that needs to be done in exploring students’ statistical reasoning when sampling and making point estimates and many important research questions exist that the above research has not addressed. There are two areas that should be the focus for future researchers. The first area is a need for research that also studies students’ quantification of the level of confidence (Ben-Zvi et al., 2011; Prodromou, 2011) when engaged in an informal statistical inference task that involved making point estimates of a population parameter within a computer-based simulation.

The estimation tasks used in the reported study could be expanded to include student expression of their level of confidence in relation to their informal inferential reasoning while sampling. This could be achieved, for example, by letting the students sample until they are confident that they have a "good" estimate (i.e., the students decide when to stop sampling), rather than instructing them to do a specified number of trials.

The second area is a need to investigate the relationship between statistical thinking and mathematical thinking in order to promote statistical thinking in relevant learning situations for students being the outcome of a balanced synthesis of ideas and information from the context sphere, statistics sphere and mathematics sphere.

References

Australian Curriculum, Assessment and Reporting Authority. (2011). *Australian Curriculum: Mathematics*. Version 1.2. Retrieved March 15, 2011, from http://www.acara.edu.au

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64-83.

Bakker, A., Kent, P., Derry, J., Noss, R., & Hoyles, C. (2008). Statistical inference at work: Statistical process control as an example. *Statistics Education Research Journal, 7*(2), 131-146.

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching of Statistics*, Salvador, Bahia, Brazil, 2-7 July, 2006. Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/*iase/publications/17/2D1_BENZ.pdf (Accessed 10 December, 2012)

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Children's emergent articulations of uncertainty while making informal statistical inferences. *ZDM Mathematics Education, 44*, 913-925. http://dx.doi.org/10.1007/s11858-012-0420-3

Ben-Zvi, D., Makar, K., Bakker, A., & Aridor, K. (2011). *Children's emergent inferential reasoning about samples in an inquiry-based environment*. Paper presented to the 7th Congress in Mathematics Education, Rzeszow, Poland, 9-13 February. Retrieved from http://www.cerme7.univ.rzeszow.pl/WG/5/CERME_BenZvi-Makar-Bakker-Aridor.pdf (Accessed 20 December, 2012)

Camtasia: Tec Smith Corporation. (2000). *Catania studio* (Version 6.0) [Computer software]. Okemos, MI: Tec Smith Corporation. Retrieved October 20, 2009, from http://www.techsmith.com/camtasia.asp

delMas, B. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenges of developing statistical literacy, reasoning and thinking* (pp. 79-95). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kobold, C., & Pollatsek, A. (2002). Data analysis as a search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259-289. http://dx.doi.org/10.2307/749741

Lane-Getaz, S. J. (2006). What is statistical thinking, and how does it develop? In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with chance and data* (pp. 273-290). Reston, VA: The National Council of Teachers of Mathematics.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The Reasoning Behind Informal Statistical Inference. *Mathematical Thinking and Learning, 13*, 152-173. http://dx.doi.org/10.1080/10986065.2011.538301

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82-105.

Makar, K., Wells, J., & Allmond, S. (2011). Is this game 1or game 2? Primary children's reasoning about samples during inquiry. In *Presentation Papers from the International Collaboration for Research on Statistical Reasoning, Thinking and Literacy (SRTL7)* (pp.17-39). Utrecht, the Netherlands: Utrecht University.

Mason, J., Burton, L., & Stacey K. (2010). *Thinking Mathematically* (2nd ed.). United Kingdom: Pearson Education.

Oxford University Press. (2012). *Oxford Dictionaries*. Retrieved 24 October, 2012, from http://oxforddictionaries.com/definition/english/thinking

Pfannkuch, M., & Wild, C. (2004) Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield

(Eds.), *The Challenges of developing statistical literacy, reasoning and thinking* (pp. 17-46). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Prodromou, T. (2011). Students' emerging inferential reasoning about samples and sampling: *In Proceedings of the Thirty Fourth Annual conference of the Mathematics. Education Research Group of Australasia* (pp. 640-648). Alice Springs, Australia. Retrieved from http://www.merga.net.au/documents/RP_PRODROMOU_MERGA34-AAMT.pdf (Accessed 10 December, 2012)

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-265. http://dx.doi.org/10.1111/j.1751-5823.1999.tb00442.x

Call for Manuscripts

International Journal of Statistics and Probability is an open-access, international, double-blind peer-reviewed journal published by the Canadian Center of Science and Education. This journal publishes research papers in all areas of statistics and probability. The journal is available in electronic form in conjunction with its print edition. All articles and issues are available for free download online.

We are seeking submissions for forthcoming issues. All manuscripts should be written in English. Manuscripts from 3000–8000 words in length are preferred. All manuscripts should be prepared in LaTeX or MS-Word format, and submitted online, or sent to: ijsp@ccsenet.org

Paper Selection and Publishing Process

a) Upon receipt of a submission, the editor sends an e-mail of confirmation to the submission's author within one to three working days. If you fail to receive this confirmation, your submission e-mail may have been missed.

b) Peer review. We use a double-blind system for peer review; both reviewers' and authors' identities remain anonymous. The paper will be reviewed by at least two experts: one editorial staff member and at least one external reviewer. The review process may take two to three weeks.

c) Notification of the result of review by e-mail.

d) If the submission is accepted, the authors revise paper and pay the publication fee.

e) After publication, the corresponding author will receive two hard copies of the journal, free of charge. If you want to keep more copies, please contact the editor before making an order.

f) A PDF version of the journal is available for download on the journal's website, free of charge.

Requirements and Copyrights

Submission of an article implies that the work described has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the authorities responsible where the work was carried out, and that, if accepted, the article will not be published elsewhere in the same form, in English or in any other language, without the written consent of the publisher. The editors reserve the right to edit or otherwise alter all contributions, but authors will receive proofs for approval before publication.

Copyrights for articles are retained by the authors, with first publication rights granted to the journal. The journal/publisher is not responsible for subsequent uses of the work. It is the author's responsibility to bring an infringement action if so desired by the author.

More Information

E-mail: ijsp@ccsenet.org

Website: www.ccsenet.org/ijsp

Paper Submission Guide: www.ccsenet.org/submission

Recruitment for Reviewers: www.ccsenet.org/reviewer

The journal is peer-reviewed
The journal is open-access to the full text
The journal is included in:

- EBSCOhost
- Gale's Academic Databases
- Google Scholar
- JournalTOCs
- PKP Open Archives Harvester
- ProQuest
- Standard Periodical Directory

International Journal of Statistics and Probability

Quarterly

Publisher	Canadian Center of Science and Education
Address	1120 Finch Avenue West, Suite 701-309, Toronto, ON., M3J 3H7, Canada
Telephone	1-416-642-2606
Fax	1-416-642-2608
E-mail	ijsp@ccsenet.org
Website	www.ccsenet.org/ijsp

ISSN 1927-7032

9 771927 703008 01>