Collection of Papers in commemoration of Prof. de Waal's 34 years of service as the Head of Department of Mathematical Statistics

Most of the recent research by Prof. Daan de Waal was devoted to extreme value theory and its applications. However, he himself has created data for future researchers in this area. Indeed, 34 years of service to the Department as its Head is a remarkable extreme value!

In this volume, his friends and colleagues present some of their latest research. The first contribution "About Daan" combines experiences, memories and biographical facts.

Contents

D.J. (Daan) de Waal

Daan de Waal attended school in the town of Boshoff in the Free State. His intention had always been to take up farming on the family farm in the district after leaving school, that is, until a teacher advised him to study for a degree in agriculture at the newly established Faculty of Agriculture at the University of the Orange Free State. It was then decided that Daan would study for the four-year BSc Agric degree , hopefully to equip him better for a successful farming career. Arriving at the university at the beginning of 1959 Daan had no idea which subjects to enrol for. On learning that his best subject at school was mathematics, the then dean of the faculty, Prof. R. Saunders, enrolled him with Biometry and Statistics as majors. Prof. Saunders was a biometrician himself and had written a textbook, "Experimental Design", with Prof. A.A. Rayner of the University of Natal, Pietermaritzburg. So now Daan was enrolled for a degree in agriculture, but with majors he had never heard of.

However, after four years of study he was hooked and decided to continue with an honours degree and finally with an MSc Agric in Biometry/Statistics, which he obtained with distinction in 1964. During his post graduate studies at the Faculty of Agriculture, Daan also lectured Biometry II and Biometry III, among others to Hennie Groeneveld who later became Professor in Biometry and Statistics at the University of Pretoria. By now the idea of a farming career had gone out of the window and in 1965 Daan was appointed as lecturer in the Department of Statistics of the UOFS, where Prof. Andries Reitsma was head with Koos Oosthuizen as senior lecturer. During this time statistics was the new buzzword in mathematical sciences and statistics departments were created and growing at universities all over the country.

Daan developed an interest in multivariate analysis, largely influenced by the 1958 textbook of T.W. Anderson, which was considered the definitive work on that subject at the time. He wanted to continue with his PhD studies in that field and was advised by Prof. Dries Reitsma to contact Prof. Cas Troskie at UCT. Cas was the newly appointed head of the Statistics Department at UCT and had obtained his PhD in multivariate analysis under Prof. H.S. Steyn at UNISA a few years earlier. Daan was Cas's first PhD student, and so in 1966 began a lifelong friendship and academic association between them. Daan completed his PhD thesis on Non-Central Multivariate Beta distributions in 1968 and was then immediately offered a Senior Lectureship at UCT, which he accepted. Another important milestone in 1968 was Daan's marriage to Verena Vermaak. After a year and a half in the Cape he returned to Bloemfontein when, at the age of 29, he was offered a chair

in the Department of Mathematical Statistics at the UOFS. When he was appointed as professor in 1970, Daan was already the author of 6 publications, two of them published in the Annals of Mathematical Statistics. He was also the supervisor of two Ph.D students, Daan Nel and Nico Crowther.

At the instigation of Prof. Norman Johnson, Daan visited during 1974-75 the University of North Carolina at Chapel Hill, where he gave a course in multivariate analysis. Thereafter the family, which by that time had grown to five, travelled to Stanford, California in an old Pontiac, a journey that took eleven days. Daan spent three months at Stanford where he met and had discussions with people like Ingram Olkin, Ted Anderson, Charles Stein, Brad Efron and Carl Morris. There he also met and started a lifelong friendship with Jim Zidek from the University of British Columbia. It is difficult to say where Daan's interest in Bayesian statistics started but it could have been during this time when the Stein estimator was causing a stir in the statistical community. He was interested in the Stein estimator and from there it was a natural step to empirical Bayes and then Bayesian thinking. During the 1980's Daan turned into a pure Bayesian, a viewpoint that was strengthened over the years through repeated visits by well known Bayesians such as Dennis Lindley, Jim Berger, Seymour Geisser, Jose Bernardo, Arnold Zellner and Jim Press.

After the untimely death of Prof. Dries Reitsma, Daan became head of department in 1977, a post he has now held for 32 years. Under his leadership the Department of Mathematical Statistics at the University of the Free State has gone from strength to strength. It now has a permanent teaching staff of 15 full time and 6 part-time lecturers, with four professors, and teaches more than 2 000 students each semester. Daan's passion for research and his ability to motivate people has instilled a culture of research in the staff and at any one time there are about 4 or 5 PhD students in the department. To date, twenty students have completed their doctorates under Daan's supervision, the first being Daan Nel in 1972, who recently retired as a Professor at the University of Stellenbosch. Five of Daan's PhD students have served as Presidents of the South African Statistical Association. Daan has published about 65 papers in statistical journals and in 1985 he was awarded the Havenga prize for mathematics by the S.A. Academy of Arts and Sciences. He has also three times received the ESKOM Excellence Award for work done on water inflow into the Gariep Dam and in 2004 he received the University of the Free State Excellence Award.

Apart from statistics, the other great passion in Daan's life is sailing. During a summer vacation in Mossel Bay in 1979 he was lying on the beach sunburned and bored. Watching

a man sailing a dinghy on the river he told Verena "That man is enjoying his holiday more than I am enjoying mine". So in 1979 he bought a dinghy which he sailed on the dams around Bloemfontein. Soon the family (now of size six) complained that the boat was too small and he bought a 26ft Elvstöm. Daan kept the boat on the Gariep Dam and in time he served as commodore of the Gariep Dam Yacht Club. In 1984, to his pride and joy, Daan graduated to a 38ft Fahr called Dahverene, an acronym formed from the names of all the members of the family. Daan enjoys nothing better than sailing over a weekend to a deserted island in the dam, with the family or some friends having a braai and spending the night on the water. However the laptop usually comes along, for when the wind is still there may be time for some statistical calculations. Daan's enthusiasm for boats and sailing is shared by his family, although we don't know how much choice Verena had about acquiring the enthusiasm! Recently both Daan's daughters were crewing for a luxury chartered cruiser out of Fort Lauderdale in the U.S, and each of his two sons has a sailing boat, although more modest than the flagship of the family.

Apart from his teaching and doctoral students Daan has been leading a SANPAD research group (South Africa – Netherlands research programme) on new models in Survival Analysis related to AIDS, in collaboration with staff from Delft Technical University and the University of Fort Hare. He has also been an active team member of a South African – Belgium research group on Extreme Value Theory where his knowledge of Multivariate Analysis and Bayesian Theory has led to new models in this field, such as the Multivariate Generalized Burr-Gamma Distribution. The Extreme Value Theory also finds important application in Daan's ongoing ESKOM project on inflows into the Gariep Dam.

Daan planned to retire in 2006 at the age of 65 as head of the Department of Mathematical Statistics and Actuarial Science, but the Dean of the Faculty of Natural and Agricultural Sciences, Prof. Herman van Schalkwyk, had such a high opinion of his teaching and research abilities that he appointed him for another three years on a contract basis. At the end of 2006 Daan received a grant of 1.8 million Rand from Eskom for research over a period of three years on water inflows into the Gariep Dam, Risk Management and Extreme Value Theory. So at present Daan is busier than ever.

Despite all his research activities and committee commitments, Daan has an intense interest in all other research being done in the department. He always has time to discuss research (or personal) problems and to suggest new ideas. He supports his staff in their endeavours but then lets them get on with these without interfering.

A short list of Daan's achievements is the following:

Personal information:

Daniël Jacobus de Waal was born in Boshof during 1941. Daan is married and has four children.

Qualifications:

1958: Matriculated from the Rooidak Hoërskool, Boshof.

1962: BSc (Agric), UOFS (Mathematical Statistics & Biometry).

1963: BSc (Agric) Hons, UOFS (Biometry & Statistics).

1964: MSc (Agric), UOFS.

1968: PhD (Mathematical Statistics), UCT.

Experience:

1964: Part-time lecturer in Biometry, UOFS.

1965 – 1968: Lecturer in Statistics, UOFS.

1969 – 1970: Senior lecturer in Mathematical Statistics, UCT.

1971 – 2006: Professor in Mathematical Statistics, UOFS.

1977 – 2006: Head of Department of Mathematical Statistics, UOFS.

2000 – 2002: Vice Dean, Natural Sciences, UOFS.

2002: Acting Dean, Natural and Agricultural Sciences, UOFS.

2007 – 2009: Head of Department of Mathematical Statistics, UFS (Contract appointment)

Other appointments and cooperation:

UFS Council member (1988 – 2006)

SA-Flemish Cooperation project as a team member (2000 – 2006).

SA-Netherlands cooperation project (SANPAD) as project leader (2000 – 2004).

ESKOM project leader (1997 – 2009)

SA Statistics Council member (2007 – 2009)

Awards:

1985: Havenga award for Mathematics of the SA Academy of Arts and Sciences.

1993: FRD evaluation "B".

1998 & 2001 & 2002: Eskom Excellence award.

2004: University of the Free State Excellence award.

Publications:

65 publications in international statistical journals and 130 technical reports.

Membership of Societies:

Member, Fellow and past President of SASA.

Elected member of the ISI.

Member, SA Academy for Arts and Science.

Member Council for Natural Scientists.

Piet Groenewald and Abrie van der Merwe

Daan de Waal: A Personal Message

It is a great privilege to communicate this message in recognition of the major contribution that Daan has made to the wellbeing and development of the subject of Statistics. I am specifically referring to the impact of his work and effort in South Africa.

As early as 1962, Daan and I were two of eight students who shared Bungalow 2 of the Reitz Bungalows at the University of the Free State. At this point Daan was in his fourth year of study and I was in my first year of study. These were wonderful and memorable years and we even devoted some time to academic issues. I still remember very well how Daan spent time with us as junior students to understand mathematics and statistics.

Towards the end of the sixties Daan was my promoter for the DSc degree at the University of the Free State, which I completed in 1972. This again was a wonderful experience doing research with Daan.

The topic that I have chosen for my paper stems directly from my thesis and the research I did with Daan. It is based on conditional distributions in a multivariate normal framework which appears in the thesis. I think Daan will remember it and hope that he still enjoys it. Unfortunately it does not contain any Bayesian ideas.

Daar bestaan vir my geen twyfel nie dat Daan sal voortgaan om so 'n positiewe invloed op sy kollegas en studente te h$\hat{e}$. Ek is ook daarvan oortuig dat hy sal voortgaan met sy navorsing en sal voortgaan om so 'n voortreflike leier in die Suid-Afrikaanse statistiese gemeenskap te wees. Ons is baie dankbaar daarvoor

Nico Crowther

Daan

I remember early 1993 and I am sitting in my office in St. Petersburg. Then the secretary of the director of my Institute comes with a fax from South Africa (!?) with invitation to visit UOVS for 6 months. It was signed: Prof. DE WAAL. This is how I saw this name for the first time. At that stage Daan was involved in a project on safety and reliability of a nuclear power station and that is how my name has attracted his attention. There was also BLOEMFONTEIN in the message and I remembered the name of the city from my childhood (the Businar's novel : Pieter Maritz the young boor from Transvaal) This seemed to be quite exciting at that stage and appeared to be the main adventure of my life.. Anyway, in a few months he was already meeting me at the airport and my nearly 6 months experience in South Africa started, and the collaboration and the friendship with Daan for all these years to come which is very important for me in my professional and personal life.

Then Daan and Verena visited us in St. Petersburg and there was a story about this visit. Those who know Daan, are aware that he is a story-teller; nice and witty stories; sometimes real sometimes from his mind, but always funny. This probably goes from a boor tradition of story-telling in small towns (as described by Herman Charles Bosman of whom I am an admirer due to my wife). And the story about him was that owing to some probably statistical reason he decided at first that August has 30 days and had informed me that they were coming on the last day of this month. Later it came to his knowledge, I assume, that August is indeed longer on one day. But the initial 'last day setting' was not replaced by the new updated information. You can imagine our anxiety (we phoned to airlines, train stations, etc) when they did not show up on the 30th and eventually had arrived only the next day in accordance with the prior estimate.

Daan is a multitalented person: a brilliant statistician; a proud husband, father and grandfather; a passionate yachtsman. He enjoys life in all diversity and this is also a talent. No doubt, it is not so easy to be a head of the department for such a long period of time. I witnessed the last 10 years as a staff member and must admit that his calm, reasonable manner of dealing with different (sometimes sensitive) matters is really remarkable.

I visited once our colleague in the US and when Daan's name was mentioned in our conversation, he said: a man with a big heart. And this, I believe, is very true.

Maxim Finkelstein

About Daan

I know Daan from our student years at the UOFS, but my first academic cooperation with him was in 1968 when he was lecturer in Statistics at the UOFS, while I was lecturer in Mathematics teaching Linear Algebra. He was working on his PhD in Multivariate Analysis at UCT under Cas Troskie and our mutual interest in matrices brought us into contact. The research he was doing on multivariate distribution theory was very interesting and when he completed his PhD, I enquired to enroll for a PhD with him as advisor when he was at UCT. During 1969 till 1971 I worked with him and could visit him on two occasions. His enthusiasm and his remarkable intuition for unsolved problems impressed me ever since.

During 1971 he was appointed as professor at UOFS and I changed my enrollment to UOFS for the rest of the study which was completed in 1972. By then I was committed to statistics and multivariate analysis and decided to become a real statistician. After two years at the Statistics Department at Stellenbosch University I returned to UOFS in 1975. This was the beginning of a very happy and rewarding academic relationship with him and other colleagues in the Department of Mathematical Statistics UOFS for the next 24 years. After the untimely death of Prof. Andries Reitsma in 1976, Daan was appointed as head of the Department of Mathematical Statistics and I succeeded Prof Reitsma. This was a great honour, particularly to be working with Daan. His style of leadership always impressed me by the trust and support he had in his colleagues and personnel. His support and assistance with research and organizational matters regarding the department and courses to be presented was easy going, cooperative and never demanding. He gave his staff ample opportunities to develop by attending conferences abroad and he assisted in many ways with this. We had the freedom to investigate different kinds of problems.

His enthusiasm for statistics and research was crowned with success with many PhD students completing, many publications and the Havenga prize for Mathematics of the SA Academy for Arts and Science awarded to him. Visitors from all disciplines in statistics visited the department at the UOFS and we were privileged to communicate and interact with them. Later he became interested in Bayesian Statistics and the analysis of rare events and made great contributions in these fields too.

I wish him a happy retirement with good health and the time to reflect and maybe pursue some of the grottos in Statistics where we enthusiastically just shoveled at the entrances. May there be many more surprises waiting!

Daan Nel

Daan as I know him

My first meeting with Daan de Waal was at the 1970 conference of the South African Statistical Association, held on the old campus of the (then) RAU. I was in the first year of my doctoral study and gave a talk on my research on quadratic forms in i.i.d. random variables and its role in goodness-of-fit. This being my first talk at a conference, I was rather nervous and when someone in the audience asked a very polite question whether my results were related to the well-known results on quadratic forms in normal variables (of which I knew nothing), the only reply I could give was a very abrupt "No!". That person happened to be Daan as I realised afterwards.

Since that first meeting with Daan, I have come to know him well and have had the pleasure of serving as external examiner to the Department over many years. I have come to admire him for the way he built their Department in Bloemfontein. Over the many years that he played a leading role, the Department produced a steady stream of research output and PhD students. Daan's own research developed over a number of areas, multivariate analysis, Bayes analysis and extreme value theory. He was instrumental in having a large number of well known academics visit their Department, with of course a spill-over benefit to other departments in South Africa.

There is of course the other side of Daan to enjoy, his sense of humour and his passion for sailing. What a joy to watch the sun set over the Gariep from the stern of his yacht.

Daan has had a long and fruitful career as academic, over the years a role model to many young statisticians. For that I thank him and wish him many more healthy and productive years.

Tertius de Wet

Prof. de Waal as my promotor

Ek glo die langste pad wat ‘n mens kan loop is ‘n gang tussen twee kantoordeure, die afstand tussen jou deur en jou promotor se deur.

Eers is hierdie pad ‘n donker pad, ‘n onseker pad, ‘n pad wat jy halfpad loop en dan omdraai, terug na jou eie kantoor om seker te maak of jy nie self die problem kan oplos nie. Dit is ‘n pad waarlangs jy 100 keer ‘n vraag oordink net vir ingeval jy dom gaan klink, of prof se tyd gaan mors.

Later verander hierdie donker pad in ‘n helder pad met ‘n lig aan die einde van die tonnel. Met ‘n kantoordeur wat vriendelik nooi om te klop en ‘n kantoor waar jy welkom voel, waar jy tuis voel. Dit is hier waar jou probleme en vrae beantwoord word, waar die onsekerheid begin minder word, waar die flou vlammetjie van “ek kan navorsing doen” al helderder begin brand.

Ten einde laaste verander hierdie pad in ‘n bekende pad, ‘n pad wat jou voete toe-oë kan loop, wat al uitgetrap is van al die besoeke. Dit is dan wanneer die langste pad ‘n kortpad word, ‘n pad van hoop vir dinge wat nog vermag kan word, omdat jy weet aan die einde van die pad is iemand wat jy kan vertrou.

Dit was vir my ‘n ongeloofike voorreg om prof. de Waal as promotor te kon hê, ‘n ekspert in statistiek en veral ekstremewaardes. Om saam met iemand te werk wat soveel kennis het om te deel was voorwaar ‘n ervaring. As promotor was prof. de Waal baie ondersteunend en positief oor my navorsing. Hy was altyd beskikbaar, maak nie saak hoe besige hy was nie, altyd maar geduldig om weer ‘n keer vir my te verduidelik. Vir my is prof de Waal verseker ‘n mentor waarna ek met die grootste respek kan opkyk. Ek hoop dat ek as studieleier in sy voetspore kan volg.

Dankie prof vir al die tyd en energie wat prof bereid is om af te staan aan prof se studente.

Prof. de Waal as my promoter

I believe that the longest path one can walk is a corridor between two office doors, the distance between your door and your promoter's door.

First this path is a dark path, an uncertain path, a path that you walk halfway and then turn back to your own office to make sure whether you can't solve the problem yourself. It is a path along which you rethink a question a 100 times afraid that you might sound stupid or waste prof's time.

Later this dark path turns into a clear path with a light at the end of the tunnel. An office door that friendly invites you to knock, an office where you are welcome and where you feel at home. It is here where your problems and questions are answered, where the uncertainty begins to fade, where the faint flame of "I can do research" starts to burn brighter.

At last this path turns into a familiar path, a worn out path from all the visits which your feet can walk eyes-closed. This is then that the longest path becomes a short path, a path of hope for things that can still be achieved, because you know at the end of the path there is someone you can trust.

It was a great privilege to have Prof. de Waal as my promoter, an expert in statistics and especially extreme values. To work with someone that has so much knowledge to share is indeed an experience. As my promoter Prof. de Waal was very supportive and positive about my research. He was always available and answered my questions patiently, even though he was busy. To me Prof. de Waal is definitely a mentor I can look up to with great respect. I hope that I will be able to follow in his footsteps as a study leader.

Thank you professor for all the time and energy you are willing to give your students.

Andrehette Verster

Control Chart for the Sample Variance Based on Its Predictive Distribution

by

A.J. van der Merwe and D. Chikobvu

ABSTRACT

This paper proposes a control chart for the sample variance. A Bayesian approach is used to incorporate parameter uncertainty based on the predictive distribution of the sample variance. When the sample size is small the rejection limit for the proposed control chart tends to be wide, so that both the mean and standard deviation of the run length are large. Therefore, not knowing the value of the population variance, σ^2, has a considerable effect on the rejection region and thus the run length.

Keywords: Bayesian procedure, Control chart, Sample variance, Predictive distribution, Run length.

1. INTRODUCTION

Statistical process control (SPC) techniques help to improve product quality by reducing the variability of a process. Such techniques allow one to monitor processes through control charts (CCs).

In a process, two classes of sources of variation are typically thought to exist: sources of variation that cannot be economically identified and removed (chance or common causes) and sources of variation that can (special or assignable causes). Control charts play an important role among SPC techniques, and are used to detect changes in a process and identify source(s) of variation with assignable causes, and thereby reduce or eliminate variability. A CC is a graphical display of the values of a quality characteristic over time. The chart typically contains control limits (CLs) derived from statistical considerations. These limits are set so that the quality characteristic is expected to fall between them with high probability if the process is stable (in-control state). If an observation falls outside the CLs, then it is suspected that some special causes of variation other than the common ones have acted in the process (Deming 1986).

There is a large literature on SPCs and, in particular, on CCs. Woodall and Montgomery (1999) and Woodall (2000) gave an overview of research issues and ideas related to this field. They pointed out that the issue of parameter estimation has received only relatively modest attention in the area of CCs.

The focus of the present paper is the formal incorporation of parameter uncertainty in the construction of control charts for the sample variance. As mentioned by Human, Chakraborti and Smit (2009) the variance chart is particularly important since an estimate of the variance is required for setting up a control chart for the mean. Thus the variance of the process must be monitored and controlled before (or even simultaneously with) attempting to monitor the mean. Similarly to Menzefricke (2002, 2007) a Bayesian approach will be used to incorporate parameter uncertainty by using the predictive distribution to construct the control chart and to

obtain the control chart limits. The literature on construction of control charts using Bayesian methods seems to be relatively sparse. For example, Woodward and Naylor (1993) developed a Bayesian method for controlling processes for the production of small numbers of items. Arnold (1990) developed an economic $\overline{X}$-chart for the joint control of the means of independent quality characteristics, and Bayarri and Garcia-Donato (2005) used a Bayesian sequential procedure to establish control limits for u-control charts.

As is now well accepted in the literature, SPC is implemented in two phases in practice, referred to as Phase I and Phase II, respectively (see for example Woodall (2000)). Phase I is also called the retrospective phase and leads to the construction of the control chart limits. The construction of accurate control limits in Phase I is critical for the monitoring of the process in Phase II. Similarly to Menzefricke (2000 and 2007) our analysis is largely concerned with Phase I. Here it is assumed that a random sample is available from a stable process whose stability is to be monitored. As mentioned by Bayarri and Garcia-Donato (2005), the natural distribution for establishing a control limit at a future time f is the predictive (marginal) distribution of the statistic (sample variance in our case) at time f. Therefore, using a Bayesian approach, the predictive distribution of the sample variance will be derived to obtain the control chart limits. Assuming that the process remains stable the predictive distribution is used to derive the distribution, mean and standard deviation of the run length.

An outline of the paper is as follows: In Section 2 the predictive distribution of the sample variance is derived and in Section 3 the distribution, mean and standard deviation of the run length for a specific example are simulated. In Section 4 we evaluate of the control chart and the conclusion is given in Section 5

.

2. PREDICTIVE DISTRIBUTION OF A FUTURE SAMPLE VARIANCE

Assume that a random sample of m independent rational subgroups, each of size $n > 1$, is available and known to have come from a stable process. Furthermore, assume that the sample is from a normal distribution with unknown mean μ and unknown variance σ^2. The data are represented as $y_{ij} \sim iidN(\mu, \sigma^2)$ where y_{ij} is the $j\,th$ observation from the $i\,th$ subgroup, $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$. Since both μ and σ^2 are unknown and no prior information is available, the conventional noninformative default prior (Jeffreys' independence prior)

$$\pi(\mu, \sigma^2) \propto \sigma^{-2} \tag{2.1}$$

will be specified for those parameters.

Combining the prior with the likelihood it follows (see, for example, Zellner 1971) that the conditional posterior distribution of μ is normal, namely

$$\mu \left| \sigma^2, y \sim N\left(\overline{y}.., \frac{\sigma^2}{nm}\right)\right. \tag{2.2}$$

In the case of the variance component, σ^2, the posterior distribution is given by

$$p\left(\sigma^2 \left| \underline{y}\right.\right) = \left(\frac{\tilde{s}}{2}\right)^{\frac{1}{2}k} \frac{1}{T\left(\frac{k}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}(k+2)} e^{-\frac{1}{2}\frac{\tilde{s}}{\sigma^2}} \qquad \sigma^2 > 0\,. \tag{2.3}$$

which is an Inverse Gamma distribution with $k = m(n-1)$ and $\tilde{s} = m(n-1)S_p^2$.

Furthermore,

$$\bar{y}_{..} = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij} \ , \qquad S_p^2 = \frac{1}{m}\sum_{i=1}^{m} s_i^2 \ ,$$

$$S_i^2 = \sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2 \ , \qquad \bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

and $\underline{y} = \left[y_{11}\ y_{12} \cdots y_{1_n}\ y_{21}\ \ y_{22} \cdots y_{2n} \cdots y_{m1}\ y_{m2} \cdots y_{mn}\right]'$.

As mentioned above, in Phase I it is assumed that a random sample is available from a stable process whose stability is to be monitored. A predictive distribution derived from a Bayesian approach can be used to obtain the control chart limits. Since the focus is on predictive distributions, we envision a future sample of n independent observations from a normal distribution, and denote the future sample variance by S_f^2.

For a given σ^2 it follows that $\frac{(n-1)s_f^2}{\sigma^2} \sim \chi_{n-1}^2$ which means that

$$f\left(s_f^2|\sigma^2\right) = \frac{v^{\frac{1}{2}}\left(\sigma^2\right)^{-\frac{1}{2}v}}{\mathrm{T}\left(\frac{v}{2}\right)2^{\frac{1}{2}v}}\left(s_f^2\right)^{\frac{1}{2}v-1} e^{-\frac{1}{2}\frac{vs_f^2}{\sigma^2}} \quad s_f^2 > 0 \text{ where } v = n-1 \tag{2.4}$$

The unconditional predictive density of S_f^2 is given by

$$f\left(s_f^2 \left| \underline{y}\right.\right) = \int_0^\infty f\left(s_f^2|\sigma^2\right)p(\sigma^2|y)d\sigma^2$$

$$= \frac{v^{\frac{1}{2}v}(\tilde{s})^{\frac{1}{2}k}\mathrm{T}\left(\frac{k+1}{2}\right)}{\mathrm{T}\left(\frac{v}{2}\right)\mathrm{T}\left(\frac{k}{2}\right)}\left(s_f^2\right)^{\frac{1}{2}v-1}\left(vs_f^2 + \tilde{s}\right)^{\frac{1}{2}(v+k)} \quad s_f^2 > 0 \tag{2.5}$$

From (2.5) it is clear that

$$S_f^2 \left| \underline{y} \sim S_p^2 F_{v,k} = S_p^2 F_{n-1}, m(n-1)\right. \tag{2.6}$$

where

$F_{n-1}, m(n-1)$ denotes a F-distribution with $n-1$ and $m(n-1)$ degrees of freedom.

The predictive distribution for S_f^2 in (2.6) can be used to obtain the control chart limits. The resulting rejection region of size α is defined as

$$\alpha = \int_{R(\alpha)} f\left(s_f^2 \left| \underline{y} \right.\right) ds_f^2.$$

Assuming that the process remains stable, this predictive distribution can also be used to derive the distribution of the "run length". Typically, a "future" sample of size n is taken repeatedly from the process, and one wants to determine the distribution of the "run length", that is the number of such samples, r, until the control chart signals for the first time. (Note that r here does not include the sample when the control chart signals). Given σ^2 and a stable process, the distribution of the run length r is geometric with parameter $\psi(\sigma^2) = \int_{R(\alpha)} f\left(s_f^2|\sigma^2\right) ds_f^2$, where $f\left(s_f^2|\sigma^2\right)$ is given in (2.4). For given σ^2, the future samples are independent of each other. The value of σ^2 is of course unknown and the uncertainty is described by its posterior distribution in (2.3), denoted by $p(\sigma^2|y)$.

The predictive distribution of the "run length" or the "average run length" can therefore easily be simulated. The first two moments of r can also be obtained by numerical integration, namely

$$E(r|y) = \int \frac{1}{\psi(\sigma^2)} p\left(\sigma^2 \left| \underline{y} \right.\right) d\sigma^2 \text{ and } E(r^2|y) = \int \frac{2-\psi(\sigma^2)}{\psi(\sigma^2)} p\left(\sigma^2 \left| \underline{y} \right.\right) d\sigma^2.$$

3. EXAMPLE

Table 3.1 displays $m = 20$ rational subgroups, each of size $n = 5$, simulated from a normal distribution; for our current purpose the mean and the variance of the normal distribution from which the samples were simulated are not mentioned because we assume that both these parameters are unknown. Also shown in Table 3.1 are the sample variances, S_i^2, for $i = 1,2, \ldots,20$. These data will be used to construct a Shewhart-type Phase I upper control chart for the variance, and also to calculate the run length for a future sample of size n taken repeatedly from the process.

Table 3.1: Data for constructing Shewhart-type Phase I upper control chart for the variance

Sample number / Time (i)	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	S_i^2
1	23.0	27.8	21.5	24.3	18.9	10.93

2	14.2	25.9	27.3	17.9	19.1	30.77
3	24.7	16.6	22.8	26.9	21.5	15.03
4	23.6	20.8	28.4	18.6	24.5	13.95
5	14.1	20.9	18.2	19.0	28.7	28.85
6	23.0	13.4	29.4	28.4	11.6	68.83
7	19.5	14.9	23.3	12.1	11.2	26.20
8	16.8	25.5	19.2	19.7	23.6	12.39
9	15.1	18.1	22.3	18.4	23.0	10.64
10	17.5	16.0	19.1	26.8	23.1	19.42
11	26.2	24.3	22.0	21.4	25.9	4.82
12	15.9	23.2	17.8	16.6	13.8	12.41
13	14.8	17.0	19.1	13.1	15.0	5.32
14	13.8	18.3	25.0	18.2	18.5	16.03
15	28.2	23.2	16.6	18.8	18.7	21.53
16	12.9	20.0	32.2	16.4	26.1	59.47
17	22.0	11.9	21.5	21.1	17.9	17.80
18	21.1	19.4	16.3	21.8	14.3	10.23
19	16.2	21.4	25.5	14.2	28.0	34.67
20	12.5	17.2	17.9	14.4	16.5	4.92

$$S_p^2 = \frac{1}{n}\sum_{i=1}^{n} S_i^2 = \frac{1}{20}(10.93 + 30.77 + \cdots + 4.92) = 21.21$$

For $\alpha = 0.01,\ n = 5,\ m = 20$ and by using the predictive distribution defined in equation (2.6) the upper control limit is given by $S_p^2 F_{n-1}, m(n-1)(\alpha) = 21.21(3.56) = 75.51$. Inspection of Table 3.1 shows that all sample variances are smaller than the upper control limit of 75.51.

Given σ^2 and a stable process the distribution of the run length r is geometric with parameter

$\psi(\sigma^2) = \int_{R(\alpha)} f\left(s_f^2|\sigma^2\right) ds_f^2$, where $f\left(s_f^2|\sigma^2\right)$ is defined in (2.4) and

$R(\alpha) = \left(s_p^2 F_{n-1,m(n-1)}(\alpha)\ ;\ \infty\right) = (75.51\ ;\ \infty)$.

Therefore, for given σ^2,

$$p\left(s_f^2 > s_p^2 F_{n-1,m(n-1)}(\alpha)\right) = P\left(\frac{\sigma^2 \chi_{n-1}^2}{n-1} > s_p^2\ F_{n-1,m(n-1)}(\alpha)\right) \quad \text{(from (2.4))}$$

$$= P\left(\frac{m(n-1)s_p^2}{\chi_{m(n-1)}^2} \quad \frac{\chi_{n-1}^2}{n-1} > s_p^2\ F_{n-1,m(n-1)}(\alpha)\right) = P\left(\chi_{n-1}^2 > \frac{1}{m}\chi_{m(n-1)}^2 F_{n-1;m(n-1)}(\alpha)\right) \quad \text{(from (2.3))}$$

$$= \psi\left(\chi_{m(n-1)}^2\right) \text{ (for given } \chi_{m(n-1)}^2) \quad (2.7)$$

By simulating ℓ values from a chi-square distribution with $m(n-1) = 80$ degrees of freedom and calculating (2.7), the distribution, mean and variance of the run length r can easily be obtained. For our example we used $\ell = 10000$.

In Figure 3.1 the distribution of the run length r is illustrated in the form of a histogram, and in Figure 3.2 the distribution of the average run length $\psi^{-1}\left(\chi^2_{m(n-1)}\right)$ is given. The means and standard deviations are also presented.

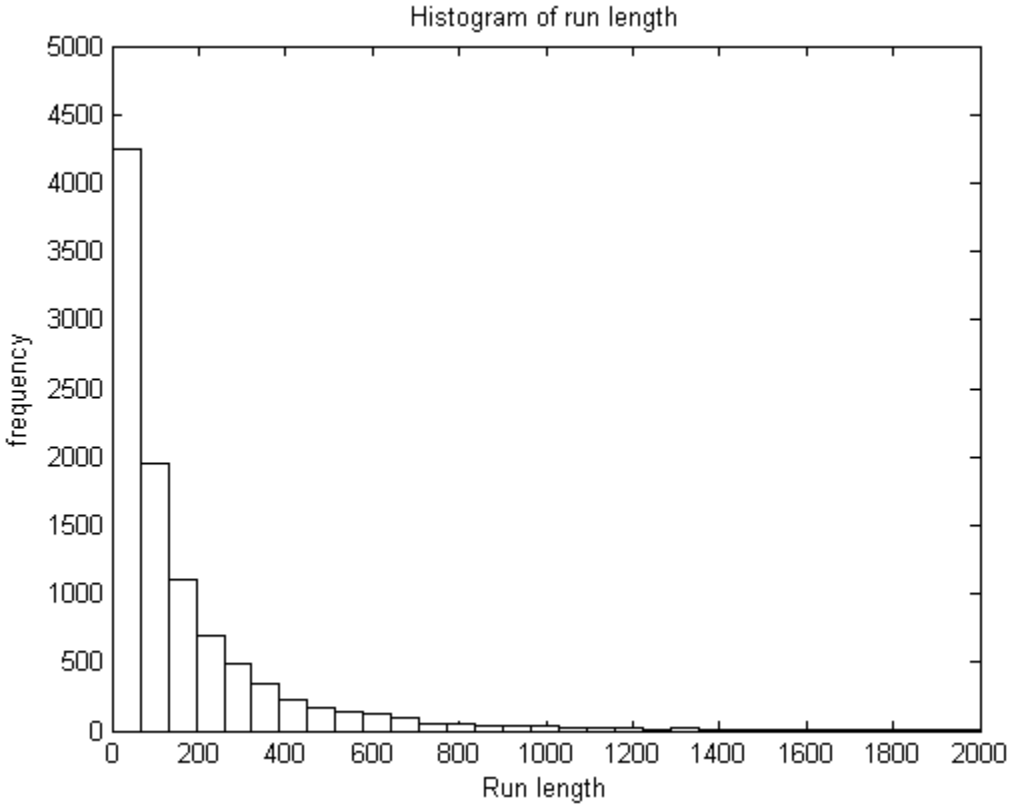


Figure 3.1 Distribution of the run length
$\boldsymbol{E\left(r\left|\underline{y}\right.\right) = 265.99}$ **and** $\boldsymbol{std\ \left(r\left|\underline{y}\right.\right) = 675.74}$

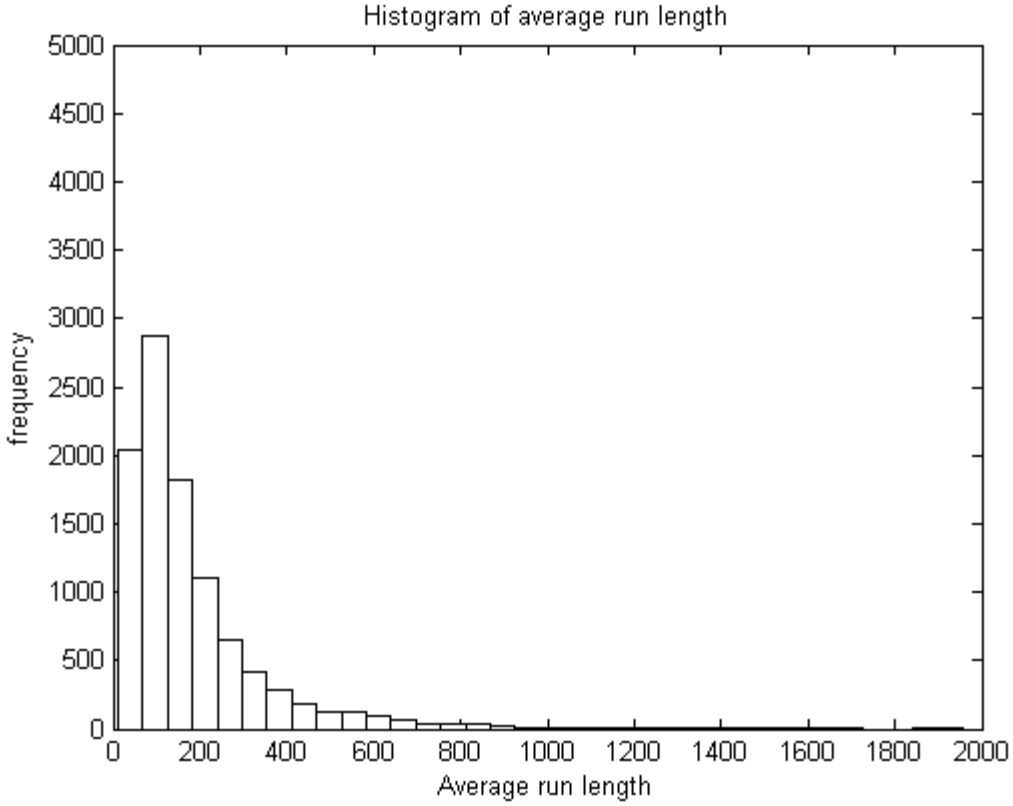


Figure 3.2: Distribution of the average run length $\psi^{-1}\chi^2_{m(n-1)}$)
$E(\psi^{-1}(\chi^2_{m(n-1)}) = 268.12$ and $Std\,(\psi^{-1}(\chi^2_{m(n-1)}) = 440.23$

The figures suggest that the means of the two distributions are the same but the standard deviations differ. This is not surprising since $\psi(\chi^2_{m(n-1)})$ is used as a parameter to simulate r. The standard deviation of r is therefore larger than that of $\psi^{-1}\chi^2_{m(n-1)})$.

4. EVALUATION OF THE CONTROL CHART

In this section we will numerically illustrate the effect of different degrees of parameter uncertainty on various aspects of the control chart. The rejection region will consist of large values for the sample variance. From equation (2.7) it is clear that without loss of generality it can be assumed that $S_p^2 = 1$. For $\alpha = 0.01$ Table 4.1 lists the mean run length and the standard deviation of the run length for $n = 5$ and selected values of m. Note that $m(n-1)$ is the posterior sample size and it thus measures the degree of parameter uncertainty regarding the sample variance. A large value of m suggests relatively little parameter uncertainty. When m is small the rejection limit for the proposed control chart tend to be wide, hence both the mean and standard deviation of the run length are large. When $m \to \infty$, there is no parameter uncertainty and the expected run length is $\frac{1}{\alpha} = \frac{1}{0.01} = 100$. It is clear that not knowing σ^2 has a very considerable effect on the rejection region and thus the run length.

Table 4.1 Expected run length and standard deviation for $n = 5$ and selected values of m

Selected values of m	$E\left(r\,\middle|\,\underline{y}\right)$	$Std\left(r\,\middle|\,\underline{y}\right)$

10	1169.9	17275.99
11	842.78	8979.70
12	662.46	5323.44
13	546.12	3478.22
14	466.03	2449.51
15	407.45	1824.06
16	365.54	1431.47
17	332.09	1158.88
18	306.68	973.91
19	285.25	833.24
20	268.76	730.73

5. CONCLUSION

In this note Bayesian methods are used to incorporate parameter uncertainty into the construction of a control chart for the sample variance based on its predictive distribution. A large sample size (large m) suggests relatively little parameter uncertainty. When m is small the rejection limit for the proposed control chart tends to be wide, hence both the man and standard deviation of the run length are large. It is therefore clear that not knowing σ^2 has a very considerable effect on the rejection region and thus the run length.

REFERENCES

Arnold, B.F. 1990. *An Economic $\overline{X}$-Chart to the Joint Control of the Means of Independent Quality Characteristic.* Zeitschrift für Operations Research, 34, 59-74.

Bayarri, M.J. and Garcia-Donato, G. 2005. *A Bayesian Sequential Look at u-Control Charts.* Technometrics 47(2): 142 – 151.

Deming, W. 1986. *Out of the crisis. Cambridge, M.A.* MIT Center for Advanced Engineering Study.

Human, S.W., Chakraborti, S., Smit, C.F. *Control Charts for Variation in Phase 1 Applications.* Submitted to Computational Statistics and Data Analysis.

Menzefricke, U. 2002. *On the Evaluation of Control Chart Limits based on Predictive Distributions.* Communications in Statistics – Theory and Methods, 31 (8): 1423-1440.

Menzefricke, U. 2007. *Control Charts for the Generalized Variance based on its Predictive Distribution.* Communications in Statistics – Theory and Methods, 36: 1031-1038.

Woodall, W.H.; Montgomery, D.C. 1999. *Research Issues and Ideas in Statistical Process Control.* Journal of Quality Technology, 31(4), 376-386.

Woodall, W.H. 2000. *Controversies and Contradictions in Statistical Process Control.* Journal of Quality Technology 32: 341-350.

Woodward, P.W.; Naylor, J.C. 1993. *An Application of Bayesian Methods in SPC.* The Statistician, 42, 461-469.

ON TERMINATING POISSON PROCESSES IN SOME SHOCK MODELS

Maxim Finkelstein
Department of Mathematical Statistics
University of the Free State, Bloemfontein,

and

Francois Marais
CSC, Cape Town, South Africa

ABSTRACT. A system subject to a point process of shocks is considered. Shocks occur in accordance with the homogeneous Poisson process. Different criteria of system failure (termination) are discussed and the corresponding probabilities of failure (accident) free performance are derived. The described analytical approach is based on deriving integral equations for each setting and solving these equations via the Laplace transform. Some approximations are analyzed and further generalizations and applications are discussed.

Keywords: Poisson process, shocks, probability of termination, Laplace transform, time of recovery

1. INTRODUCTION

Consider first, a general point process $\{T_n\}; T_0 = 0, T_{n+1} > T_n, n = 0,1,2,\ldots$, where T_n is time to the nth arrival of an event with the corresponding cumulative distribution function (Cdf) $F^{(n)}(t)$. Let G be a geometric variable with parameter θ (independent of $\{T_n\}_{n\geq 0}$) and denote by T a random variable with the following Cdf:

$$F(t,\theta) = \theta \sum_{k=1}^{\infty} \overline{\theta}^{k-1} F^{(k)}(t), \tag{1}$$

where $\overline{\theta} = 1-\theta$.

A natural reliability interpretation of model (1) is via the *stochastic point process of shocks*. Let T be a random time to failure (termination) of a system subject to a point process of shocks [1]. We interpret the term "shock" in a very broad sense as some instantaneous, potentially harmful event. Assume for simplicity that a shock is the only cause of failure. It means that a system is 'absolutely reliable' in the absence of shocks. Assume also that each shock independently of the previous history leads to a system failure with probability θ and is survived with probability $\overline{\theta} = 1-\theta$. This procedure defines the terminating point process, whereas the corresponding survival probability of our system (reliability) in $(0,t)$ is $P(t,\theta) \equiv 1 - F(t,\theta)$.

Obtaining probability $P(t,\theta)$ is an important problem in various reliability and safety assessment applications. As described, the shock can have an interpretation of a 'killing' event. Alternatively, a shock process can have a meaning of a process of demands for service, whereas the survival probability is the probability that all demands are serviced. Another interpretation is when a repairable system described by the alternating renewal process should be just available at each instance of demand. In this

case the survival probability has a meaning of multiple availability [2], which is a generalization of the conventional availability.

It is clear that a general relationship (1) does not allow for explicit results that can be used in practice and therefore, assumptions on the form of the point process should be made. Two specific point processes are mostly used in reliability applications, i.e., the Poisson process and the renewal process. This paper is devoted to the case of the Poisson process of shocks. Some results for the terminating renewal processes can be found, e.g., in reference [3]. Also see the discussion in Section 5.

Consider the Poisson process of shocks with rate λ. In this case the survival probability can be easily explicitly obtained [4]:

$$\begin{aligned}\Pr[T \geq t] = P(t,\theta) &= \sum_{0}^{\infty} (\bar{\theta})^k \exp\{-\lambda t\} \frac{(\lambda t)^k}{k!} \\ &= \exp\{-\theta \lambda t\}. \end{aligned} \tag{2}$$

It follows from equation (2) that the corresponding failure rate, which describes the lifetime of our system T, is given by a simple and meaningful relationship:

$$\lambda(t) = \theta \lambda. \tag{3}$$

Thus, the rate of the underlying Poisson process λ is decreased by the factor $\theta \leq 1$. Equation (3) describes an operation of thinning of the Poisson process for this specific case [5].

The main methodological aim of this paper is to show how the method of integral equations can be effectively applied to obtaining probability $P(t,\theta)$ in various settings. In order to illustrate this claim in the simplest way, let us derive (2) using the corresponding integral equation and the subsequent Laplace transform. It is easy to see that the following equation with respect to $P(t,\theta)$ holds:

$$P(t,\theta) = e^{-\lambda t} + \int_{0}^{t} \lambda e^{-\lambda x} \bar{\theta} P(t-x,\theta) dx. \tag{4}$$

Indeed, the first term on the right hand side is the probability that there are no shocks in $[0,t)$ and the integrand defines the probability that the first shock has occurred in $[x, x+dx)$, was survived and then the system has survived in $[x,t)$. Due to the properties of the homogeneous Poisson process, the probability of the latter event is $P(t-x,\theta)$.

Applying the Laplace transform to both sides of equation (4) results in

$$\tilde{P}(s,\theta) = \frac{1}{s+\lambda} + \frac{\lambda \bar{\theta}}{s+\lambda} \tilde{P}(s,\theta) \Rightarrow \tilde{P}(s,\theta) = \frac{1}{s+\lambda\theta}, \tag{5}$$

where $\tilde{P}(s,\theta)$ denotes the Laplace transform of $P(t,\theta)$. The corresponding inversion, as in (2), results in $\exp\{-\theta \lambda t\}$.

Note that this solution is due to the fact that the Laplace transform can be nicely obtained for the case of the (homogeneous) Poisson process. For the nonhomogeneous Poisson process (NHPP) with rate $\lambda(t)$, similar to (2), direct summation gives

$$P(t,\theta)=\sum_{0}^{\infty}(\overline{\theta})^{k}\exp\left\{-\int_{0}^{t}\lambda(u)du\right\}\frac{\left\{\int_{0}^{t}\lambda(u)du\right\}^{k}}{k!}=\exp\left\{-\theta\int_{0}^{t}\lambda(u)du\right\}.$$

Generalization of integral equation (4) to the case of NHHP is not so straightforward but in principle can be performed (see Section 5). However, it is difficult to apply the 'explicit' Laplace transform in this case. Therefore, our models will be considered only for homogeneous Poisson processes of shocks.

We will discuss three models for obtaining survival probability in different settings. Other settings and generalizations can be studied as well. The main analytical tool allowing for explicit solutions in all considered situations is the method of integral equations and the subsequent application of the Laplace transform. Section 2 is devoted to the case when probability of termination depends on the quality of the repairable system's performance at the time of a shock arrival. This is a natural assumption, as resistance to a shock often depends on the state of a system. Section 3 deals with additional source of termination of the process when the shocks are too close and a system has not enough time to recover after the previous shock. In Section 4, two different types of shocks are considered. Two consecutive shocks of the first kind can kill a system, but if there is a shock of another kind between them, the system survives.

2. POBABILITY OF TERMINATION DEPENDS ON A SYSTEM'S STATE

Consider a repairable system with instantaneous, perfect repair that starts functioning at $t=0$. Let its lifetime be described by the Cdf $F(t)$, which is a governing distribution for the corresponding renewal process with the renewal density function to be denoted by $h(t)$. Assume that the quality of performance of our system is characterized by some deterministic for simplicity function of performance $Q(t)$ to be called the quality function [6]. The considered approach can be generalized to the case of a random $Q(t)$. It is often a decreasing function of time, and this assumption is quite natural for degrading systems. In applications, the function $Q(t)$ can describe some key parameter of a system, *e.g.*, the decreasing in time accuracy of the information measuring system or effectiveness (productivity) of some production process. As repair is perfect, the quality function is also restored to its initial value $Q(0)$. It is clear that the quality function of our system at time t is now random and equal to $Q(Y)$, where Y is a random time since the last (before t) repair.

The system is subject to the Poisson process of shocks with rate λ. As previously, each shock can terminate the performance of the repairable system and we are interested in obtaining the survival probability $P(t,\theta)$. Note that the repaired failure of the system does not terminate the process and only a shock can result in termination. Assume that the probability of termination depends on the system's quality at the time of a shock, which is a reasonable assumption, i.e., the larger the value of quality, the smaller the probability of termination. Let the first shock arrive before the first failure of the system. Denote by $\theta^{*}(Q(t))$ the corresponding probability of termination in this case. Now we are able to obtain $\theta(t)$ -the probability of termination of the operating system *by the first shock* at time instant t (and not necessarily before the first fail-

ure of the system). Using the standard 'renewal-type reasoning' [7], the following relationship for $\theta(t)$ can be derived:

$$\theta(t) = \theta^*(Q(t))\overline{F}(t) + \int_0^t h(x)\overline{F}(t-x)\theta^*(Q(t-x))dx, \tag{6}$$

where $\overline{F}(t) \equiv 1 - F(t)$. Indeed, the first term on the right-hand side of equation (6) gives the probability of termination during the first cycle of the renewal process, whereas $h(x)\overline{F}(t-x)dx$ defines the probability that the last failure of the system before t had occurred in $[x, x+dx)$ and therefore, the corresponding probability of termination at t is equal to $\theta^*(Q(t-x))$.

Thus, the probability of termination under the first shock $\theta(t)$, which is now time-dependent, has been derived. Assume that the survival of a shock also means an instantaneous perfect repair of the system (the 'repaired shock' is survived, the 'non-repaired' results in the termination). Therefore, the instants of survived shocks can be also considered as the renewal points for the system. Having this in mind, we can now proceed with obtaining the survival probability $P(t,\theta)$. Using the similar reasoning as when deriving equation (4)

$$P(t,\theta) = e^{-\lambda t} + \int_0^t \lambda e^{-\lambda x}\overline{\theta}(x)P(t-x,\theta)dx, \tag{7}$$

where $\overline{\theta}(x) \equiv 1 - \theta(x)$.

Applying the Laplace transform to equation (7), similar to (5):

$$\begin{aligned} &\tilde{P}(s,\theta) = \frac{1}{s+\lambda} + \lambda\tilde{\overline{\theta}}(s+\lambda)\tilde{P}(s,\theta) \\ &\Rightarrow \tilde{P}(s,\theta) = \frac{1}{(s+\lambda)(1-\lambda\tilde{\overline{\theta}}(s+\lambda))}. \end{aligned} \tag{8}$$

Given the functions $F(t)$ and $\theta^*(Q(t))$, equations (6) and (8) can be solved numerically, but we can still proceed with the Laplace transforms under an additional assumption that the underlying distribution is exponential, i.e., $F(t) = 1 - \exp\{-ht\}$. In this case $h(x) = h$ and the Laplace transform of equation (6) results in

$$\tilde{\theta}(s) = \tilde{\theta}^*(s+h)\left(1 + \frac{h}{s}\right), \tag{9}$$

where $\tilde{\theta}^*(s) = \int_0^\infty e^{-sx}\theta^*(Q(x))dx$ denotes the Laplace transform of the function $\theta^*(Q(t))$. Substituting (9) into (8) and taking into account that $\tilde{\overline{\theta}}(s) = (1/s) - \tilde{\theta}(s)$

$$\tilde{P}(s,\theta) = \frac{1}{s + \lambda\theta^*(s+h+\lambda)(s+h+\lambda)}. \tag{10}$$

It is easy to see that when $\theta(t) = \theta^*(Q(t)) = \theta$ and $h = 0$, equation (10) reduces to the simplest case (5).

To proceed further with inversion, we must make some assumptions on the form of the function $\theta^*(Q(t))$. Let $\theta^*(Q(t)) = 1 - \exp\{-\alpha t\}, \alpha \geq 0$, which is a reasonable assumption (as probability of termination increases as $Q(t)$ decreases with t) allowing for a simple Laplace transform. Then

$$\tilde{P}(s,\theta) = \frac{s + h + \lambda + \alpha}{s^2 + s(\lambda + h + \alpha) + \alpha\lambda} \tag{11}$$

and the inversion gives

$$P(t,\theta) = \frac{s_1 + \lambda + \alpha}{s_1 - s_2}\exp\{s_1 t\} - \frac{s_2 + \lambda + \alpha}{s_1 - s_2}\exp\{s_2 t\}, \tag{12}$$

where

$$s_{1,2} = \frac{-(h + \lambda + \alpha) \pm \sqrt{(h + \lambda + \alpha)^2 - 4\lambda\alpha}}{2}.$$

An important specific case is when the system is absolutely reliable ($h = 0$) but is characterized by the quality function $Q(t)$. Then $s_1 = -\lambda, s_2 = -\alpha; \alpha \neq \lambda$ and

$$P(t,\theta) = \frac{\lambda}{\lambda - \alpha}\exp\{-\alpha t\} - \frac{\alpha}{\lambda - \alpha}\exp\{-\lambda t\}. \tag{13}$$

If, for instance, $\theta^*(Q(t)) = 1$, which means that $\alpha \to \infty$, then, as expected, $P(t,\theta) = \exp\{-\lambda t\}$ (the probability that there are no shocks in $[0,t)$. On the contrary, if $\alpha = 0$, which means that $\theta^*(Q(t)) = 0$, the survival probability is equal to 1. Another marginal cases are defined by the value of the rate λ. If $\lambda = 0$, then again, as expected, $P(t,\theta) = 1$. On the other hand, it follows from (13) that as $t \to \infty$,

$$P(t,\theta) \to \exp\{-\alpha t\}, \tag{14}$$

which can be confusing at first sight, as one would expect that as the rate of a shock process tends to infinity, the probability of survival in $[0,t)$ should tend to 0, but this is not the case as the function $\theta^*(Q(t)) = 1 - \exp\{-\alpha t\}$, is close to 0 for small t and each survived shock is the renewal point for our system. Therefore, as the number of shocks increases, due to the properties of exponential function, relationship (14) holds.

3. TERMINATION WITH RECOVERY TIME

In the previous sections, the only source of termination was an immediate effect of a shock. Consider now another setting that can be often encountered in practical reliability and safety assessments problems. Let, similar to Section 1, each shock from the Poisson process with rate λ terminate the process with probability θ and be survived with probability $\bar{\theta} = 1 - \theta$. Assume now that termination can also occur when consecutive shocks are 'too close', which means that the system did not recover from the consequences of a previous shock. Therefore, the time for recovering should be taken into account. It is natural to assume that it is a random variable τ with the Cdf $R(t)$ (different values of damage need different time of recovering and this fact is described by $R(t)$). Thus, if the shock occurs while the system still did not recover from

the previous one, it terminates the process. It is the simplest criterion of termination of this kind. Other criterions can be also considered. As previously, we want to derive $P(t,\theta,R)$ -the probability of survival in $[0,t)$.

First, assume that a shock had occurred at $t=0$ and has been survived. Denote the probability of survival under this condition by $P^*(t,\theta,R)$. Similar to (4) and (7), the corresponding supplementary integral equation is

$$P^*(t,\theta,R) = e^{-\lambda t} + \int_0^t \lambda e^{-\lambda x}\overline{\theta}R(x)P^*(t-x,\theta,R)dx, \tag{14}$$

where the multiplier $R(x)$ in the integrand is the probability that the recovery time after the first shock at $t=0$ (and before the next one at $t=x$) is sufficient.

Applying the Laplace transform to both sides of (14) results in the following relationship for the Laplace transform of $P^*(t,\theta,R)$:

$$\tilde{P}^*(s,\theta,R) = \frac{1}{(s+\lambda)(1-\lambda\overline{\theta}\tilde{R}(s+\lambda))}, \tag{15}$$

where $\tilde{R}(s)$ is the Laplace transform of th Cdf $R(t)$.

Using probability $P^*(t,\theta,R)$ we can derive now the following integral equation with respect to $P(t,\theta,R)$:

$$P(t,\theta,R) = e^{-\lambda t} + \int_0^t \lambda e^{-\lambda x}\overline{\theta}P^*(t-x,\theta,R)dx \tag{16}$$

Indeed, as previously, the first term on the right hand side of this equation is the probability of shocks absence in $[0,t)$, $\lambda e^{-\lambda x}\overline{\theta}dx$ is the probability that the first shock has occurred (and was survived) in $[x,x+dx)$. Finally, $P^*(t-x,\theta,R)$ is the probability that the system survives in $[x,t)$.

We can obtain $P(t,\theta,R)$, applying the Laplace transform to both sides of (16), i.e,

$$\tilde{P}(s,\theta,R) = \frac{1}{s+\lambda} + \frac{\lambda\overline{\theta}}{s+\lambda}\tilde{P}^*(s,\theta,R),$$

where $\tilde{P}^*(s,\theta,R)$ is defined by (15). This gives the general solution of the problem under the stated assumptions in terms of Laplace transforms. In order to be able to invert $\tilde{P}(s,\theta,R)$, assume that the Cdf $R(t)$ is exponential, i.e., $R(t)=1-\exp\{-\gamma t\}, \gamma>0$. Performing simple algebraic transformations:

$$\tilde{P}(s,\theta,R) = \frac{s+2\lambda+\gamma-\theta\lambda}{s^2+s(\gamma+2\lambda)+\lambda^2+\gamma\lambda\theta}. \tag{17}$$

Inversion of (17) gives

$$P(t,\theta,R) = \frac{s_1+\gamma+2\lambda-\theta\lambda}{s_1-s_2}\exp\{s_1 t\} - \frac{s_2+\gamma+2\lambda-\theta\lambda}{s_1-s_2}\exp\{s_2 t\}, \tag{18}$$

where

$$s_{1,2} = \frac{-(\gamma + 2\lambda) \pm \sqrt{(\gamma + 2\lambda)^2 - 4(\lambda^2 + \gamma\lambda\theta)}}{2}.$$

Equation (18) gives an exact solution for $P(t,\theta,R)$. In applications it is convenient to use simple approximate formulas. Consider the following assumption:

$$\frac{1}{\lambda} >> \bar{\tau} \equiv \int_0^\infty (1 - R(x))dx, \tag{19}$$

where $\bar{\tau}$ denotes the mean time of recovery.

Relationship (19) means that the mean inter-arrival time in the shock process is much larger than the mean time of recovery, and this is often the case in practice. In the study of repairable systems, the similar case is usually called the *fast repair* condition. Using this assumption, similar to (3), the equivalent rate of termination for our process for $\lambda\bar{\tau} \to 0$, $\lambda t >> 1$ can be written as

$$\lambda(t) = B\lambda(1 + o(1)), \tag{20}$$

where B is the probability of termination for the *occurred shock* due to two causes, i.e., the termination immediately after the shock and the termination when the next shock occurs before the recovery is completed. Therefore, for sufficiently large t ($t >> \bar{\tau}$) the integration in the following integral can be performed to ∞ and the approximate value of B is

$$B = \theta + (1 - \theta)\int_0^\infty \lambda e^{-\lambda x}(1 - R(x))dx,$$

Assuming, as previously, $R(t) = 1 - \exp\{-\gamma t\}$, $\gamma > 0$ gives

$$B = \frac{\lambda + \theta\gamma}{\lambda + \gamma}.$$

Finally, the fast repair approximation for the survival probability is

$$P(t,\theta,R) \approx \exp\left\{-\frac{\lambda + \theta\gamma}{\lambda + \gamma}t\right\}. \tag{21}$$

It can be easily seen that when $\gamma \to \infty$ (instant recovery), relationship (21) reduces to equation (2). Note that approximate relation (20) is derived for all shocks except the first one (for which $B = \theta$), but the condition $\lambda t >> 1$ (large expected number of shocks in $[0,t)$) ensures that the error in (21) due to this cause is sufficiently small. The accuracy of the fast repair approximation (21) with respect to the time of recovery can be analyzed similar to reference [2].

4. TWO TYPES OF SHOCKS

Assume now that there are two types of shocks. As in the previous section, potentially harmful shocks (to be called *red* shocks) result in termination of the process when they are 'too close', i.e., when the time between two consecutive red shocks is smaller

then a recovery time with the Cdf $R(t)$. Therefore, in this case a system does not have enough time to recover from the consequences of the previous red shock. Assume for simplicity that the probability of immediate termination on red shock's occurrence is equal to 0 ($\theta = 0$). The model can be easily generalized to this case as well. On the other hand, our system is subject to the process of 'good' (*blue*) shocks. If the blue shock follows the red shock, termination cannot happen no matter how soon the next red shock will occur. Therefore, the blue shock can be considered as a kind of additional recovery action.

Denote by λ and β the rates of the independent Poisson processes of red and blue shocks respectively. First, assume that the first red shock has already occurred at $t = 0$. An integral equation for the probability of survival in $[0,t)$, $P^*(t,\beta,R)$ for this case is as follows:

$$P^*(t,\beta,R) = e^{-\lambda t} + \int_0^t \beta e^{-\beta x} e^{-\lambda x} \int_0^{t-x} \lambda e^{-\lambda y} P^*(t-x-y,\beta,R)\,dy\,dx$$

$$+ \int_0^t e^{-\beta x} \lambda e^{-\lambda x} R(x) P^*(t-x,\beta,R)\,dx\,, \qquad (22)$$

where

- The first term on the right hand side is the probability that there are no other red shocks in $[0,t)$;
- $\beta e^{-\beta x} e^{-\lambda x} dx$ is the probability that a blue shock occurs in $[x, x+dx)$ and no red shocks occur in $(0,x)$;
- $\lambda e^{-\lambda y} dy$ is the probability that the second red shock occurs in $[x+y, x+y+dy)$;
- $P^*(t-x-y,\beta,R)$ is the probability that the system survives in $[x+y,t)$ given the red shock has occurred at time $x+y$;
- $e^{-\beta x} \lambda e^{-\lambda x} dx$ is the probability that there is one red shock (the second) in $(0,t)$ and no blue shocks in this interval of time;
- $R(x)$ is the probability that the recovery time x is sufficient and therefore the second red shock does not terminate the process;
- $P^*(t-x,\beta,R)$ is the probability that the system survives in $[x,t)$ given the red shock has occurred at time x.

Using $P^*(t,\beta,R)$ that can be obtained from equation (22) we can now construct an integral equation with respect to $P(t,\beta,R)$-the probability of survival without assuming occurrence of the red shock at $t = 0$. Similar to (16)

$$P(t,\beta,R) = e^{-\lambda t} + \int_0^t \lambda e^{-\lambda x} P^*(t-x,\beta,R)\,dx\,. \qquad (23)$$

Applying the Laplace transform to equation (22) results in

$$\tilde{P}^*(s,\beta,R) = \frac{s+\beta+\lambda}{(s+\beta+\lambda)(s+\lambda) - \beta\lambda - \lambda(s+\beta+\lambda)(s+\lambda)\tilde{R}(s+\beta+\lambda)}\,. \qquad (24)$$

Applying the Laplace transform to equation (23):

$$\tilde{P}(s,\beta,R)=\frac{1}{s+\lambda}+\frac{\lambda}{s+\lambda}\tilde{P}^{*}(s,\beta,R)\,. \tag{25}$$

This equation gives a general solution of the problem under the stated assumptions in terms of Laplace transforms. In order to be able to invert $\tilde{P}(s,\beta,R)$, as in the previous section, assume that the Cdf $R(t)$ is exponential: $R(t)=1-\exp\{-\gamma t\}$, $\gamma>0$. Performing simple algebraic transformations

$$\tilde{P}(s,\beta,R)=\frac{s+\gamma+\beta+2\lambda}{s^2+s(\gamma+\beta+2\lambda)+\lambda^2}\,. \tag{26}$$

Inversion of (26) gives

$$P(t,\beta,R)=\frac{s_1+\gamma+\beta+2\lambda}{s_1-s_2}\exp\{s_1 t\}-\frac{s_2+\gamma+\beta+2\lambda}{s_1-s_2}\exp\{s_2 t\}, \tag{27}$$

where

$$s_{1,2}=\frac{-(\gamma+2\lambda+\beta)\pm\sqrt{(\gamma+\beta)^2+4\lambda(\gamma+\beta)}}{2}\,.$$

When γ=0, there is no recovery time and the process is terminated when two consecutive red shocks occur. In this case equation (27) reduces to relationship obtained in reference [8].

Equation (27) gives an exact solution for $P(t,\theta,R)$. Similar to Section 3, it can be simplified under certain assumptions. Assume that the fast repair condition (19) holds. The first red shock cannot terminate the process. The probability that the subsequent shock can result in termination is

$$B=\int_0^t \lambda e^{-\lambda x}\int_0^{t-x}\lambda e^{-\lambda y}e^{-\beta y}(1-R(y))dydx\,. \tag{28}$$

For the exponentially distributed time of recovery:

$$B=\frac{\lambda}{\lambda+\beta+\gamma}-\frac{\lambda}{\beta+\gamma}e^{-\lambda t}+\frac{\lambda^2}{(\lambda+\beta+\gamma)(\beta+\gamma)}e^{-(\lambda+\beta+\gamma)t}$$

For sufficiently large t, $B\approx\lambda/\lambda+\beta+\lambda$ and this approximate value can be used for subsequent shocks as well. Therefore, relationship

$$P(t,\theta,R)\approx\exp\left\{-\frac{\lambda^2}{\lambda+\beta+\gamma}t\right\}.$$

is the fast repair approximation in this case.

5. DISCUSSION

The method of integral equations, which is applied to deriving the survival probability for different shock models is an effective tool for obtaining probabilities of interest in

situations where the object under consideration has renewal points. As the considered process of shocks is the homogeneous Poisson process, each shock (under some additional assumptions) constitutes these renewal points. When a shock process is NHPP, there are no renewal points, but the integral equations can be usually also derived. For illustration, consider the corresponding generalization of equation (4). Denote by $P(t-x,x,\theta)$ the survival probability in $[x,t)$, $x<t$ for the 'remaining shock process' that started at $t=0$ and was not terminated by the first shock at time x. Note that this probability depends now not only on $x-t$ as in the homogeneous case but on x as well. Equation (4) is modified now to

$$P(t,\theta)=\exp\left\{-\int_0^t \lambda(u)du\right\}+\int_0^t \lambda(x)\exp\left\{-\int_0^x \lambda(u)du\right\}\bar{\theta}P(t-x,\theta)dx$$

It can be easily seen by substitution that

$$P(t-x,x,\theta)=\exp\left\{-\theta\int_x^t \lambda(u)du\right\},\ 0\le x,t$$

is the solution of this equation.

One can formally write integral equations for other models considered in this paper and the NHPP process of shocks, but their solutions should be obtained numerically as the explicit inversions of the corresponding Laplace transforms are not possible.

If shocks are described by the renewal process with the governing distribution $F(t)$ and the corresponding probability density function $f(t)$, the method of integral equations can be also obviously applied as in this case we also have 'pure renewal points'. For instance, the simplest equation (4) turns in this case into

$$P(t,\theta)=(1-F(t)+\int_0^t f(x)\bar{\theta}P(t-x,\theta)dx\,.$$

Applying the Laplace transform gives

$$\tilde{P}(s,\theta)=\frac{1-\tilde{f}(s)}{s(1-\bar{\theta}\tilde{f}(s))},$$

which is formally a solution to our problem in terms of the Laplace transform. Note that for given $F(t)$ it can be inverted usually only numerically. This is similar to the reasoning used for describing the Laplace transforms for standard renewal equations in the renewal theory [9].

Another generalization of (4) (and subsequent models) is to the case when $\theta(t)$ is a time-dependent probability. It is well-known that the probability of survival for the NHPP of shocks in this case is given by the following relationship:

$$P(t,\theta(t))=\exp\left\{-\int_0^t \theta(u)\lambda(u)du\right\},$$

which is an analogue of the Brown-Proschan model in the theory of imperfect (minimal) repair [10].

References

1. Barlow R. and Proschan F. (1975). *Statistical Theory of Reliability and Life Testing Probability Models,* Holt, Rinehart and Winston. 1975.
2. Finkelstein M.S. and Zarudnij V.I. (2202) Laplace transform methods and fast repair approximations for multiple availability and its generalizations, *IEEE Transactions on Reliability*, 51, 168-177.
3. Finkelstein M.S. (2003). Simple bounds for terminating Poisson and renewal processes, *Journal of Statistical Planning and Inference*, 113, 541-548.
4. Thompson W.A. (1988). *Point Process Models with Applications to Safety and Reliability*, Chapman and Hall.
5. Cox D.R. and Isham V. (1984). *Point Processes*, Chapman and Hall, London.
6. Finkelstein M.S. (2008). *Failure Rate Modeling for Reliability and Risk,* Springer.
7. Ross S.M. (1996). *Stochastic Processes*. John Wiley.
8. Venter J.P. (2007)…….. University of the Free State, M.Sc thesis.
9. Beichelt F.E. and Fatti L.P. (2002). *Stochastic Processes and their Applications*, Taylor and Francis.
10. Block H.W, Borges W, Savits T.H. (1985). Age dependent minimal repair, *Journal of Applied Probability*, 22, 370-386.

Investigating approximations and parameter estimation of the Multivariate Generalized Burr-Gamma

A Verster and DJ de Waal

Department Mathematical Statistics and Actuarial Science

University of the Free State

Bloemfontein

ABSTRACT

The MGBG is considered for modelling multivariate data especially containing extreme values. This article is an extension and comparison to an earlier article by Beirlant et al. (2000). Special attention is given to the approximations of the expected value and covariance of the MGBG and to the estimation of parameters. The Kolmogorov-Smirnov is considered for estimating some of the parameters.

KEYWORDS: MGBG, Extreme Values, Estimation, Approximation, Kolmogorov-Smirnov, QQ-plot.

1. INTRODUCTION

The multivariate Generalized Burr-Gamma (MGBG) distribution (Beirlant et al., 2000) generalizes the Burr-Gamma distribution (Beirlant et al., 2002) to a multivariate distribution which is fairly flexible to fit to multivariate data containing extremes on all or some of the variables. In this paper an alternative approach to Beirlant *et al*. (2000) is taken to estimate the parameters of the MGBG. In Section 2 properties of the MGBG are given and in Section 3 asymptotic formulae are derived for $E(\underline{X})$ and $Var(\underline{X})$. These results are tested for a few simulated data sets in Section 4. In Section 5 we discuss our alternative approach to estimate the parameters and then the procedure is applied to a simulated data set in Section 6. Section 7 shows the difference in the estimated parameter values between the approach in Beirlant et al. (2000) and our approach.

2. THE MULTIVARIATE GENERALIZED BURR-GAMMA DISTRIBUTION

The Multivariate Generalized Burr-Gamma distribution (MGBG) is an extension of the univariate generalized Burr-Gamma family of distributions and it allows the modelling of multivariate data that includes extreme values.

A random vector $\underline{X} = \left(X_1,...X_p\right)'$ is MGBG($\underline{k},\underline{\mu},\Sigma,\underline{\xi}$) distributed if the joint density function is given by the following equation:

$$f(\underline{x}) = \frac{1}{|\Sigma|^{\frac{1}{2}}} \prod_{i=1}^{p} \frac{\sqrt{\psi'(k_i)}}{\Gamma(k_i)\, x_i} \exp\left(-v_{\xi_i}(\underline{x})\right) v_{\xi_i}(\underline{x})^{k_i - 1} \left(1+\xi_i v_i(\underline{x})\right)^{-1} v_i(\underline{x}),\ 1+\xi_i v_i(\underline{x}) > 0$$

$$= \prod_{i=1}^{p} \Gamma\left(k_i, v_{\xi_i}(\underline{x})\right)$$

(1)

where

$$V_{\xi_i} = \frac{1}{\xi_i}\log\left(1+\xi_i V_i\right) \sim \text{GAM}\left(k_i,1\right),\ \text{independent for } i = 1,...,p$$

$$V_i = \exp\left\{\psi(k_i) - \sqrt{\psi'(k_i)}\,\Sigma_{(i)}^{\frac{-1}{2}}\left(\underline{Y}-\underline{\mu}\right)\right\},$$

$$\underline{Y} = \left(Y_1,...,Y_p\right)',\ Y_i = -\log\left(X_i\right),$$

$$\psi(k_i) = \frac{\partial}{\partial k_i}\log\Gamma(k_i) \text{ and } \psi'(k_i) = \frac{\partial}{\partial k_i}\psi(k_i).$$

$\psi(k_i)$ and $\psi'(k_i)$ represent the digamma and trigamma functions respectively. For dimension p, $\Sigma_{(i)}^{\frac{-1}{2}}$ is the i^{th} row of the symmetric square root matrix Σ^{-1}, where Σ is a symmetric positive definite $p \times p$ matrix. $\underline{k} = \left(k_1,...,k_p\right)$ denotes a positive vector of shape parameters, $\underline{\mu} = \left(\mu_1,...,\mu_p\right)$ denotes a vector of location parameters and $\underline{\xi} = \left(\xi_1,...,\xi_p\right)$ denotes a vector of extreme value indices. The parameter space is defined as

$\Omega = \left\{k_i > 0, -\infty < \mu_i < \infty, \Sigma > 0, -\infty < \xi_i < \infty\right\},\ i = 1,...,p$ (Beirlant *et al*. 2000, p. 113).

Remark: If $\xi_i = 0$ for $i = 1,...,p$, then $\mu_i = E(Y_i)$ and $\Sigma = Cov(Y)$. The implication of the remark is that $\underline{\mu}$ and Σ can be estimated from the data by deleting the extreme observations exceeding certain thresholds $t_i, i = 1,...,p$. This implies that $\underline{\mu}$ and Σ are not the mean and covariance of $\underline{X}$. The first question that we address is; what is $E(\underline{X})$ and $Cov(\underline{X})$?

3. APPROXIMATIONS FOR $E(X)$, $VAR(X)$ AND $COV(X)$

The approximated expected value of $\underline{X}$ is

$$E(X_i) \approx \exp\left(-\mu_{Y_i}\right) + \frac{1}{2}\sigma_{Y_i}^2 \exp\left(-\mu_{Y_i}\right) \tag{2}$$

where $\mu_{Y_i} = E(Y_i) = -\sum_{(i)}^{\frac{1}{2}} D_\psi^{-1} E\begin{bmatrix}\log V_1 \\ . \\ . \\ . \\ \log V_p\end{bmatrix} + \sum_{(i)}^{\frac{1}{2}} D_\psi^{-1}\begin{bmatrix}\psi(k_1) \\ . \\ . \\ . \\ \psi(k_p)\end{bmatrix} + \mu_i$

, $D_\psi = \text{diag}\left(\sqrt{\psi'(k_1)}, \sqrt{\psi'(k_2)}, ..., \sqrt{\psi'(k_p)}\right)$ and $\sigma_{Y_i}^2$ is the i^{th} diagonal element of $Cov(\underline{Y}, \underline{Y}') = \sum^{\frac{1}{2}} D_\psi^{-1} Cov(\log V, \log V') D_\psi^{-1} \sum^{\frac{1}{2}}$.

This approximation is proven in Appendix A.1.

The approximated variance of $\underline{X}$ is

$$Var(X_i) \approx k_i \left[\exp\left(\sum_i^{\frac{1}{2}} D_\psi^{-1} \begin{pmatrix} \log\left(\frac{e^{\xi_1 V_{\xi_1}} - 1}{\xi_1}\right) - \psi(k_1) \\ . \\ . \\ . \\ \log\left(\frac{e^{\xi_p V_{\xi_p}} - 1}{\xi_p}\right) - \psi(k_p) \end{pmatrix} - \mu_i \right) \left(\left[\sum_i^{\frac{1}{2}} D_\psi^{-1} \left(\frac{\xi_j}{e^{\xi_j k_j} - 1} \right) e^{\xi_j k_j} \right] \right) \right]^2. \tag{3}$$

This approximation is proven in Appendix A.2.

The approximated covariance between X_i and $X_{i'}^*$ is

$$Cov(X_i, X_{i'}) = a_i Cov\left(V_{\xi_i}, V_{\xi_{i'}}\right)[(a]_{i'})' \quad (4)$$

where $a_i = exp\left[\sum_i^{\frac{1}{2}} D_\psi^{-1} \begin{pmatrix} log\left(\frac{e^{\xi_1 k_1}-1}{\xi_1}\right) - \psi(k_1) \\ \vdots \\ log\left(\frac{e^{\xi_p k_p}-1}{\xi_p}\right) - \psi(k_p) \end{pmatrix} - \mu_i \right] \sum_i^{\frac{1}{2}} D_\psi^{-1} \frac{\xi_i e^{\xi_i k_i}}{e^{\xi_i k_i}-1}$ and $Cov\left(V_{\underline{\xi}}, V_{\underline{\xi}}'\right)$ is a diagonal matrix with $Var\left(V_{\xi_i}\right) = k_i$ on the diagonal.

This approximation is proven in Appendix A.3.

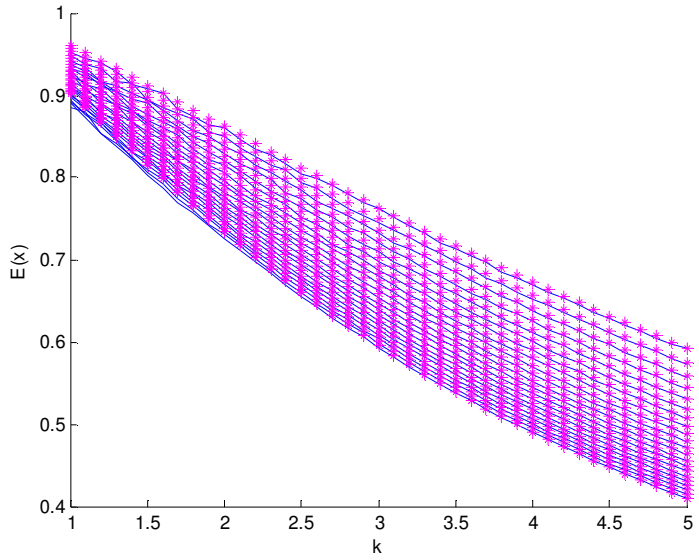
From (2), (3) and (4) it is clear that $E(\underline{X})$, $Var(\underline{X})$ and $Cov(\underline{X})$ are complicated functions of the parameters. To estimate them we propose the estimation of the parameters and substitute the estimates in these functions. In the next section we will investigate these estimations through simulations.

4. INVESTIGATING THE APPROXIMATIONS

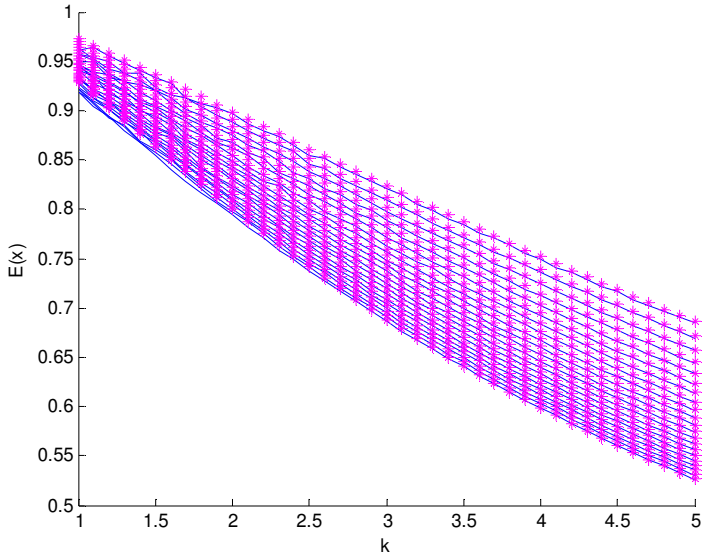
To investigate the appropriateness of the approximations for $E(\underline{X})$ and $Var(\underline{X})$ in (2) and (3), three dimensional data sets were simulated from a MGBG distribution. After simulating a data set of size $n = 500$ the approximated mean and variance are compared to the estimated mean and variance of $\underline{X}$. This is illustrated in the following figures. In Figures 1 and 2 the estimated mean and variance is indicated by the solid line and the approximated mean and variance is indicated by '*' for each variable. For the simulations $\mu = [0\ 0\ 0]$ and $\Sigma = \begin{bmatrix} 0.014 & 0.004 & 0.007 \\ 0.004 & 0.007 & 0.004 \\ 0.007 & 0.004 & 0.013 \end{bmatrix}$ are assumed. Figures 1 and 2 indicates that for $0 \leq k \leq 5$ and $-3 \leq \xi \leq -1$ the approximated mean and variance of $\underline{X}$ are rather close to the estimated mean and variance of $\underline{X}$ for each dimension.

Figure 1 Estimated $E(\underline{X})$ (indicated by -) and the approximated $E(\underline{X})$ (indicated by *) for the three different dimensions plotted against *k*.

(a) Variable 1



(b) Variable 2



(c) Variable 3

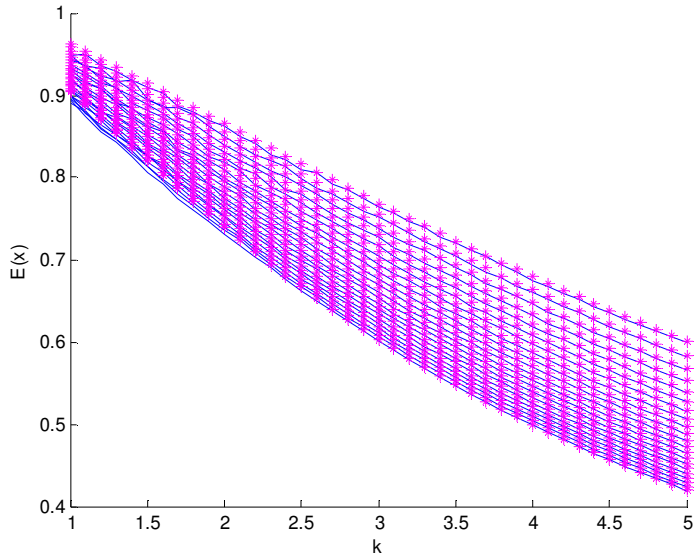
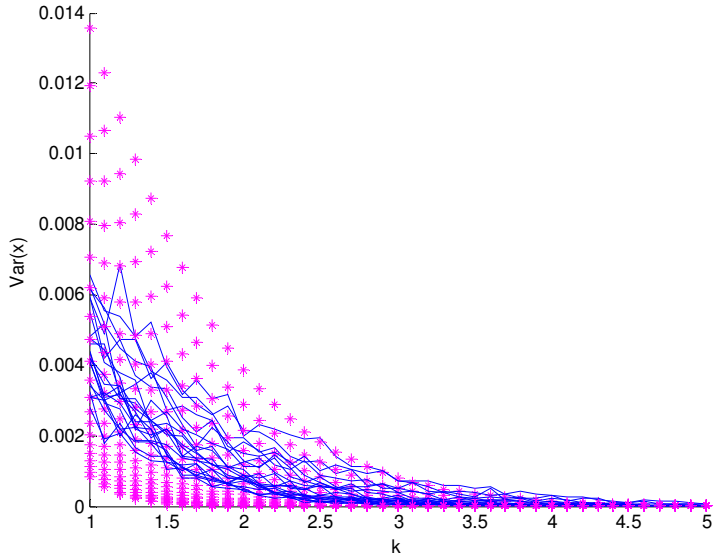
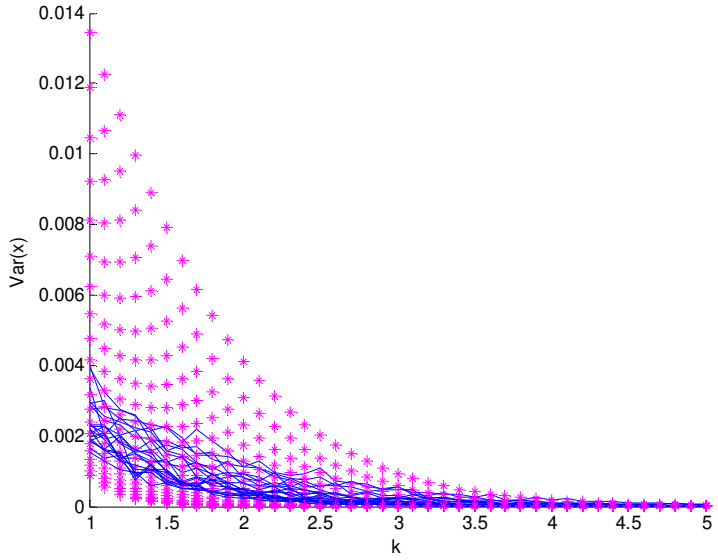


Figure 2 Estimated $Var(\underline{X})$ (indicated by -) vs. the approximated $Var(\underline{X})$ (indicated by *) for the three dimensions.

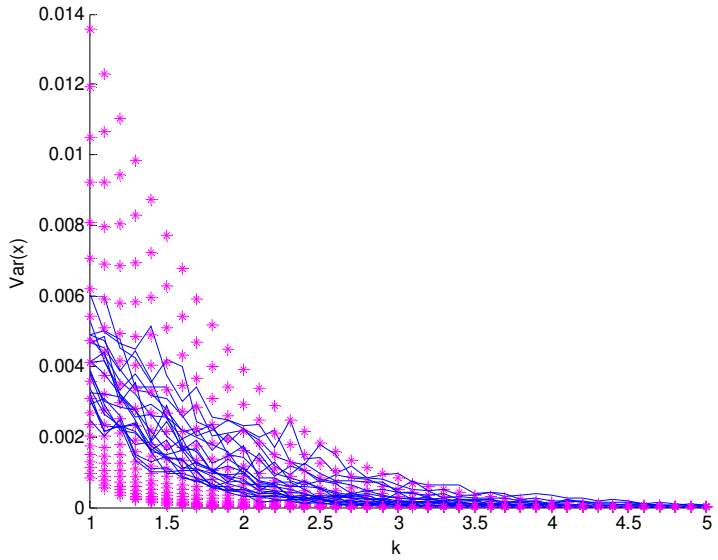
(a) Variable 1



(b) Variable 2



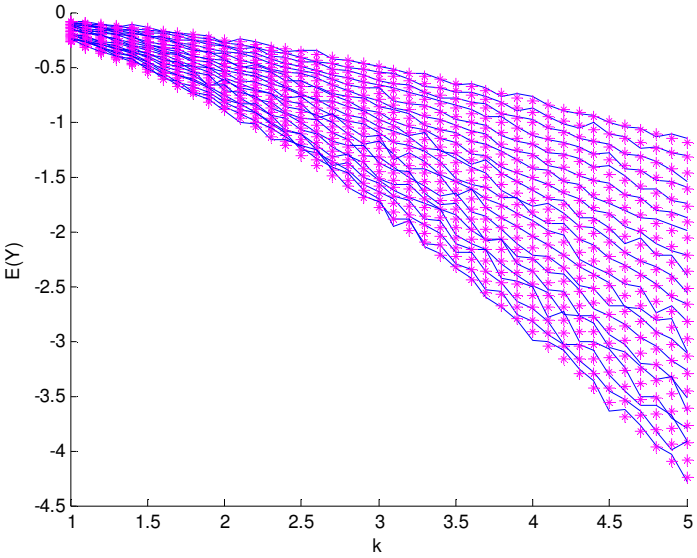
(c) Variable 3



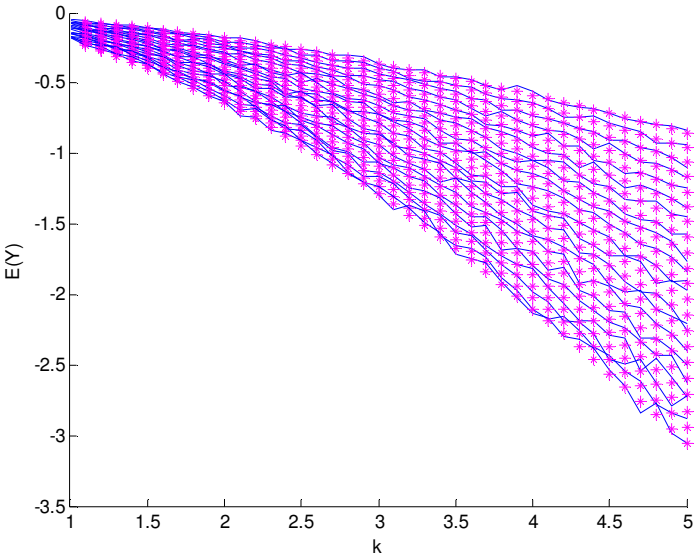
Similar simulations can be done where the estimated $E(\underline{Y})$ is compared to the approximated $E(\underline{Y})$ (Given in the Appendix). The estimated mean and variance is indicated by the solid line and the approximated mean and variance is indicated by '*'. Figures 3 and 4 show the comparison for $1 \leq k \leq 5$ and $1 \leq \xi \leq 3$. From the figures it can be seen that the approximated mean and variance of $\underline{Y}$ are close to the estimated mean and variance of $\underline{Y}$ for each dimension.

Figure 3 Estimated $E(\underline{Y})$ (indicated by -) vs. the approximated $E(\underline{Y})$ (indicated by *) for the three dimensions plotted against *k*.

(a) Variable 1



(b) Variable 2



(c) Variable 3

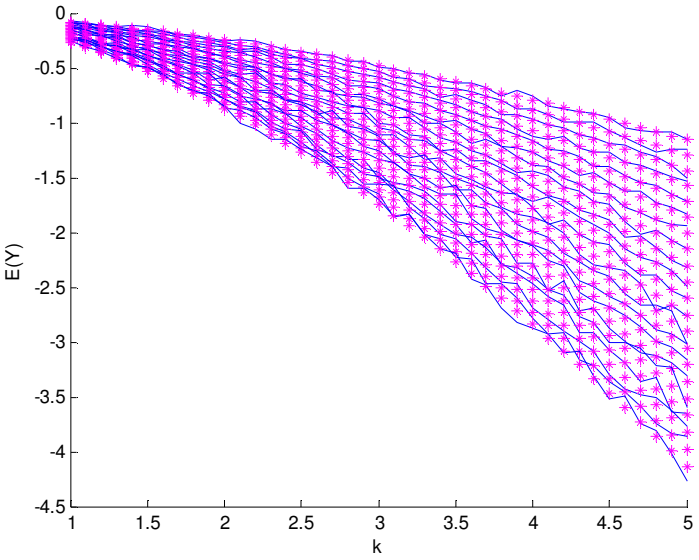
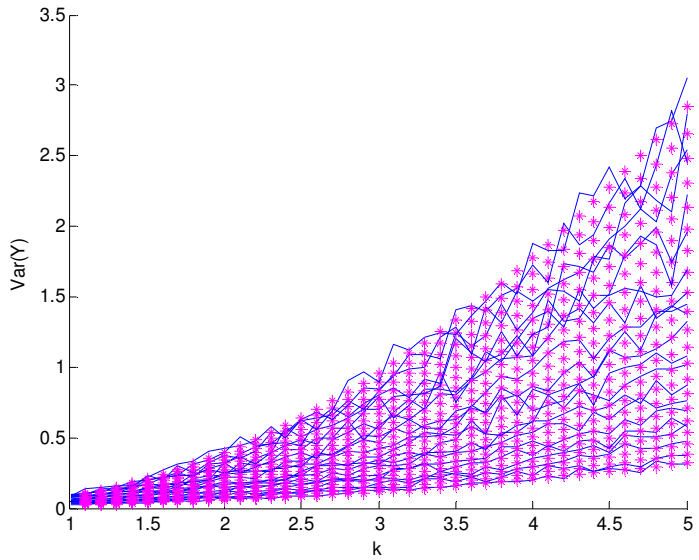
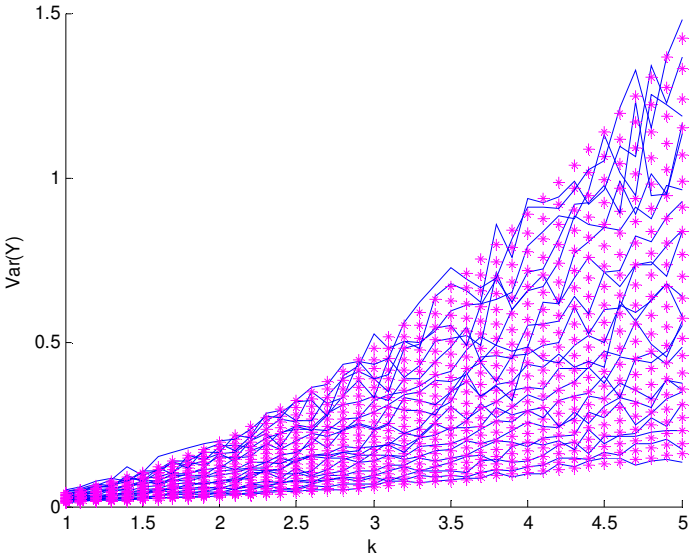


Figure 4 Estimated $Var(Y)$ (indicated by -) vs. the approximated $Var(Y)$ (indicated by *) for the three dimensions.

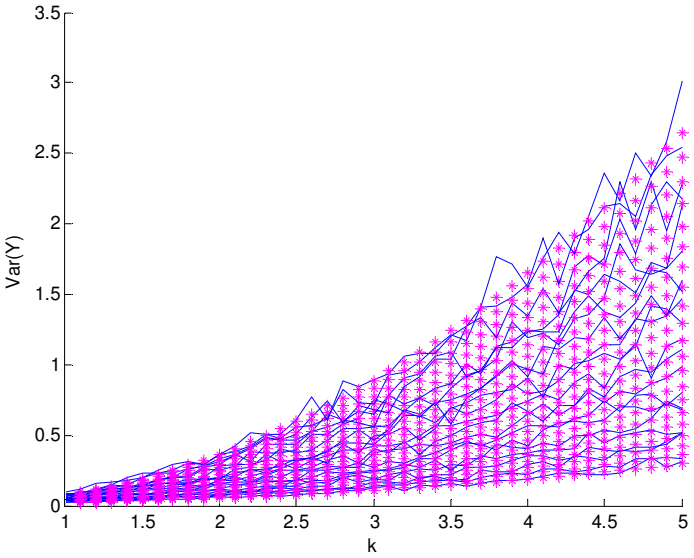
(a) Variable 1



(b) Variable 2



(c) Variable 3



For $k \geq 1$ and $\xi \geq 1$ the approximated mean and variance of $\underline{Y}$ are close to the estimated mean and variance of $\underline{Y}$. As k increases the estimated mean and variance becomes more volatile.

5. ESTIMATION OF PARAMETERS $\underline{k}$ AND $\underline{\xi}$ THROUGH THE KS MEASURE

This section discusses the estimation of the four MGBG parameters. The approach for estimating the parameters $\underline{\mu}$ and Σ are the same as in the article of Beirlant *et al*. $\underline{\mu}$ and Σ are estimated by first "trimming" the data, thus ignoring the extreme values, and then using the method of moments to estimate $\underline{\mu}$ and Σ from the data below a threshold. The method of moments are given by the following equations

$$\underline{\hat{\mu}} = \sum_{j=1}^{n_t} \left(y_{1j}, ..., y_{pj} \right)' / n_t$$

(5)

and

$$\hat{\Sigma} = \left\{ \sum_{j=1}^{n_t} \left(y_{ij}, ..., y_{pj} \right)' \left(y_{ij}, ..., y_{pj} \right) - \underline{\hat{\mu}}\underline{\hat{\mu}}' \right\} / n_t, i = 1, ..., p; j = 1, ..., n_t$$

(6)

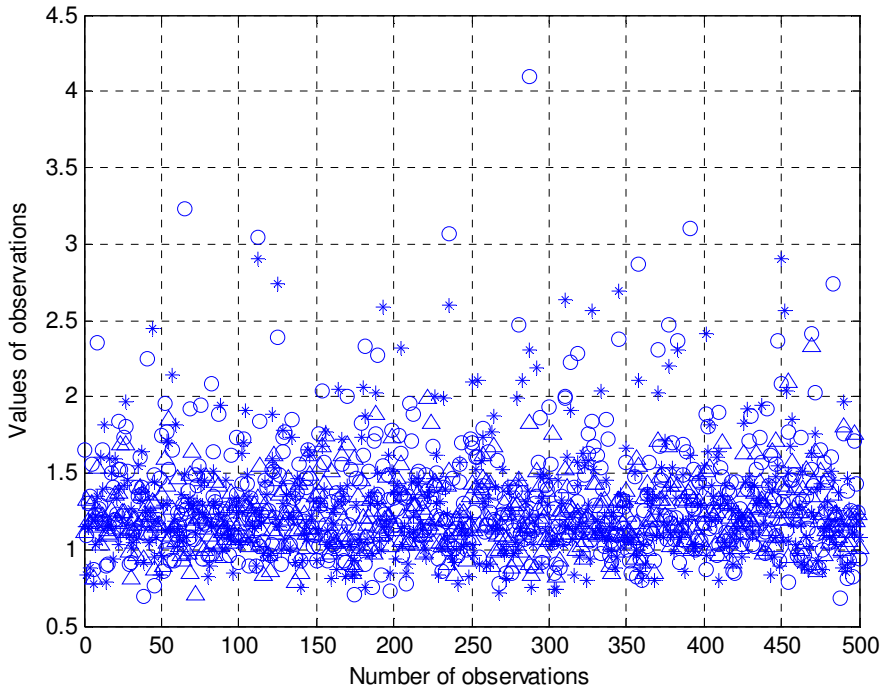
where n_t denotes the number of observations below the threshold t. Thresholds are chosen for each dimension by obtaining the 75th upper quartiles as shown in Beirlant *et al*. (2000).

The Kolmogorov-Smirnov measure, $KS = \max|F_n - F|$ (Conover, 1980) is then used to estimate values for k and ξ where F_n denotes the empirical cdf and F the fitted cdf. Since it is known that $V \sim GAM(k, 1)$ the Kolmogorov-Smirnov measure calculates the maximum absolute difference between the empirical Gamma function and the cumulative Gamma function for different values of k and ξ. With the KS measure one can see how well the model fits the data. The minimum value of the different maximum KS measure values will indicate the best fit.

6. SIMULATED DATA SETS

This section illustrates the estimation process as discussed in Section 5. $n = 500$ values $x_{ij}, i = 1,2,3; j = 1, ..., n$ were simulated from a MGBG with the following set of parameters, $\mu = [0\ 0\ 0], \Sigma = \begin{bmatrix} 0.014 & 0.004 & 0.007 \\ 0.004 & 0.007 & 0.004 \\ 0.007 & 0.004 & 0.013 \end{bmatrix}, k = [1.7\ 2.2\ 2.1]$ and $\xi = [1.2\ 0.9\ 1.0]$. Figure 5 shows the simulated data values of $X_{ij}, i = 1, ...p;\ j = 1, ..., n$. Let $Y_{ij} = -logX_{ij}, i = 1, ..., p;\ j = 1, ..., n$.

Figure 5 Simulated value of $\underline{X}$, * indicates variable 1, Δ indicates variable 2 and o indicates variable 3.



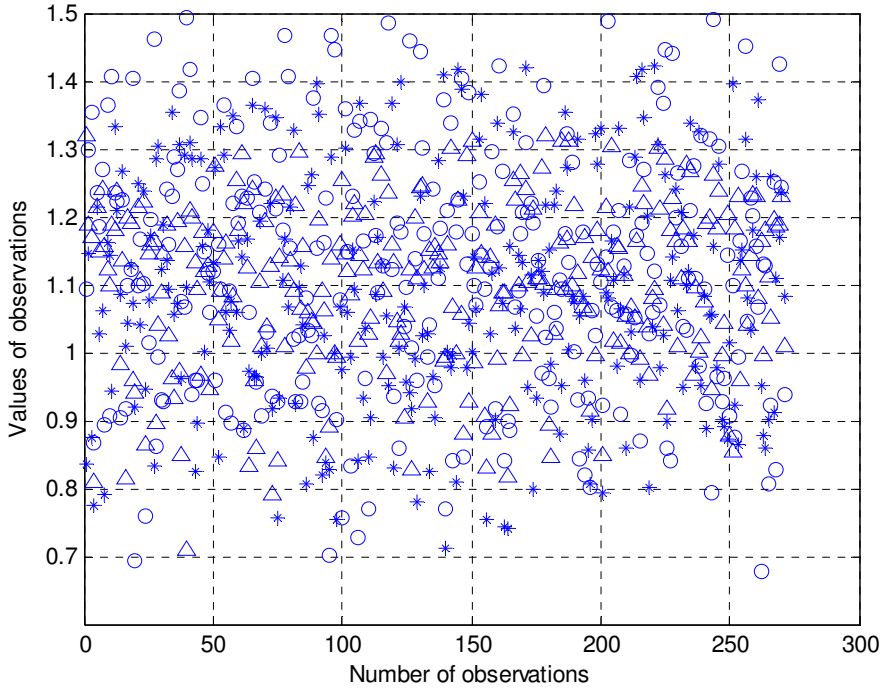
From Figure 5 it is evident that extreme values occur in the data set. The thresholds are now determined by obtaining the 75th upper quartile in every dimension.

Figure 6 shows the simulated $\underline{X}$ values below the thresholds. The mean and the variance of $\underline{Y}$ is now calculated for the data below the threshold as follows:

$\underline{\hat{\mu}} = [-0.0731 \; -0.0897 \; -0.0997]$ and $\hat{\Sigma} = \begin{bmatrix} 0.0258 & 0.0043 & 0.0105 \\ 0.0043 & 0.0135 & 0.0054 \\ 0.0105 & 0.0054 & 0.0281 \end{bmatrix}$.

Figure 6 Simulated value of $\underline{X}$ below the thresholds, * indicates variable 1, Δ indicates variable 2 and o indicates variable 3.

.



For different values of $1 < k < 5$ and $1 < \xi < 1.9$ the KS measure is calculated. The estimates of $\underline{k}$ and $\underline{\xi}$ are the values of $\underline{k}$ and $\underline{\xi}$ that gives the minimum KS measure value. The estimated parameter values are shown in Table 1 together with the minimum KS measure value. Table 1 also includes the estimated parameter values when the threshold is chosen as the 75th percentile.

Table 1 Estimated vs true parameter values.

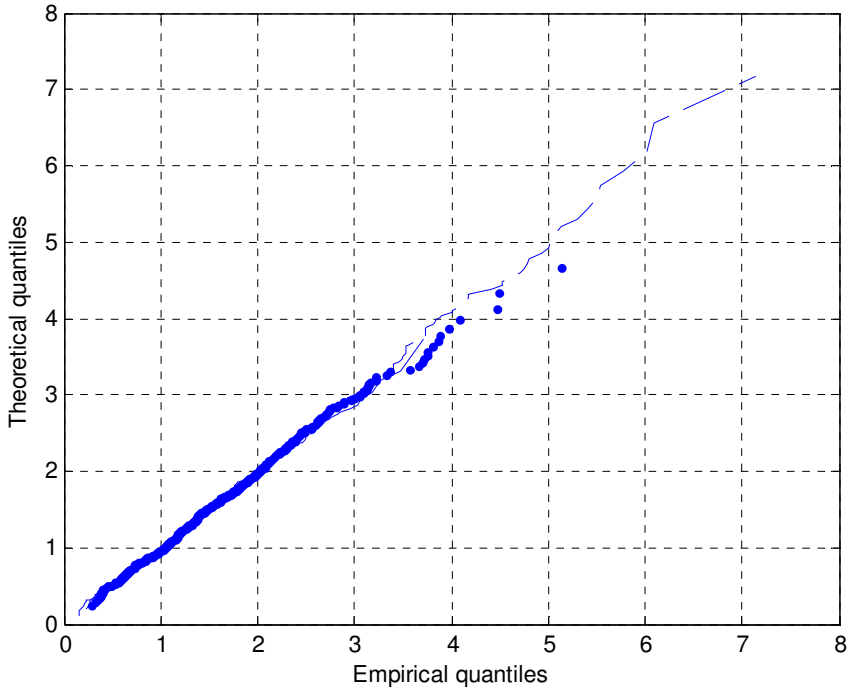
	True parameter value	**Estimated parameter values for t = 75th quartiles**
μ	[0 0 0]	[−0.0731 − 0.0897 − 0.0997]
Σ	$\begin{bmatrix} 0.014 & 0.004 & 0.007 \\ 0.004 & 0.007 & 0.004 \\ 0.007 & 0.004 & 0.013 \end{bmatrix}$	$\begin{bmatrix} 0.0258 & 0.0043 & 0.0105 \\ 0.0043 & 0.0135 & 0.0054 \\ 0.0105 & 0.0054 & 0.0281 \end{bmatrix}$
k	[1.7 2.2 2.1]	[2.4 2.3 2.6]
ξ	[1.2 0.9 1.0]	[1.7 1 1.2]

The minimum KS values for each dimension are $[0.0179\ 0.0236\ 0.0196]$, these values are all significant at the 5% level.

From Table 1 it is clear that the estimated parameter values are relatively close to the true parameter values.

Figure 7 shows the QQ-plots for each dimension for the estimated parameters in Table 1. A QQ-plot gives an indication of the goodness of fit of the MGBG to the simulated data. If the QQ-plot follows more or less a straight line it indicates a good fit, as is the case in Figure 7.

Figure 7 QQ-plots for each dimension, dimension 1 is indicated by the straight line, dimension 2 by the dashed line and dimension 3 by the dots.



7. APPLICATIONS TO REAL DATA

In this section a real data set, the maximum monthly wind speed from March 1993 to December 1998, recorded in three stations in the Cape Town area, namely Cape Town harbour (HB), Cape Town airport (AP) and Robben Island (RI), are modelled with the MGBG. The same data set was considered previously by Beirlant *et al*. (2000) and the results are compared with our results. The data is given in the Appendix A4. Let $\underline{X} = (X_{HB}, X_{AP}, X_{RI})$ denote the wind speed in knots. The same thresholds were used as in the article by Beirlant *et al*. (2000), t = (60, 55, 45). When considering only the data below the threshold $\underline{\mu}$ and Σ were estimated as

$$\underline{\hat{\mu}} = \begin{bmatrix} -3.8770 \\ -3.8873 \\ -3.5899 \end{bmatrix} \text{ and } \hat{\Sigma} = \begin{bmatrix} 0.0135 & 0.0038 & 0.0068 \\ 0.0038 & 0.0070 & 0.0039 \\ 0.0068 & 0.0039 & 0.0124 \end{bmatrix}$$

which is the same as the results in Beirlant *et al*. (2000).

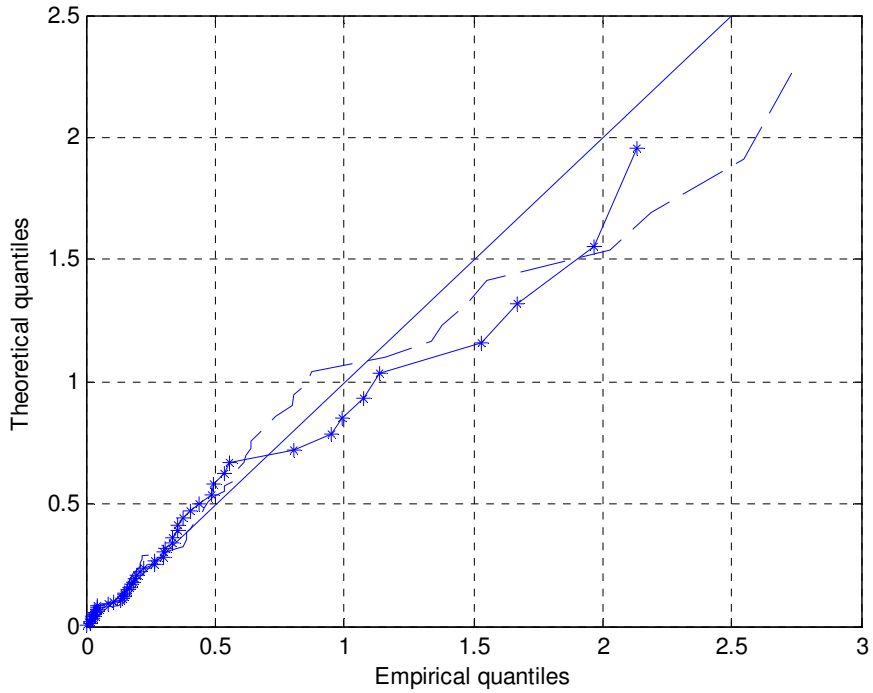
$\underline{\xi}$ and $\underline{k}$ were estimated simultaneously by using the Kolmogorov-Smirnov approach for different values of $0.3 < k < 3$ and $0.2 < \xi < 7$ and the following estimates were obtained

$\hat{\underline{\xi}} = [6.2 \;\; 0.5 \;\; 1]$ and $\hat{\underline{k}} = [0.4 \;\; 0.8 \;\; 0.5]$.

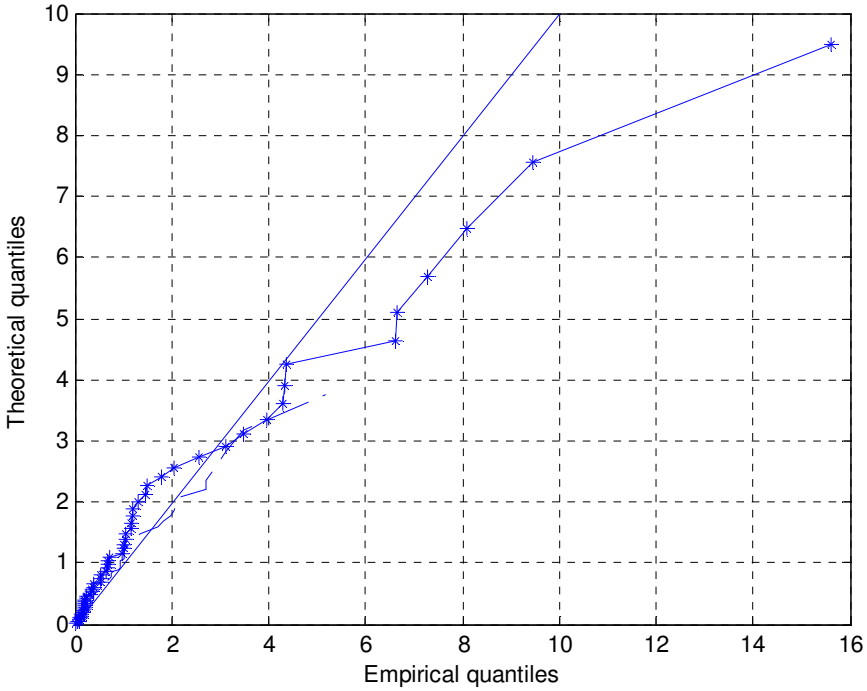
The estimates of $\underline{\xi}$ and $\underline{k}$ differs considerable from the estimates in Beirlant *et al*. (2000) where $\hat{\underline{\xi}} = [2.8867 \;\; 1.3645 \;\; 0.3997]$ and $\hat{\underline{k}} = [0.8 \;\; 1.2 \;\; 1.3]$. QQ-plots with our parameters and the parameters of Beirlant *et al*. (2000) are shown in Figure 8 for the different dimensions.

Figure 8 QQ-plots for each dimension, (*-) represents the QQ-plots with our parameters and the (- -) represents the QQ-plots with the parameters of Beirlant *et al*. (2000) and the solid line represents the 45^{o} line.

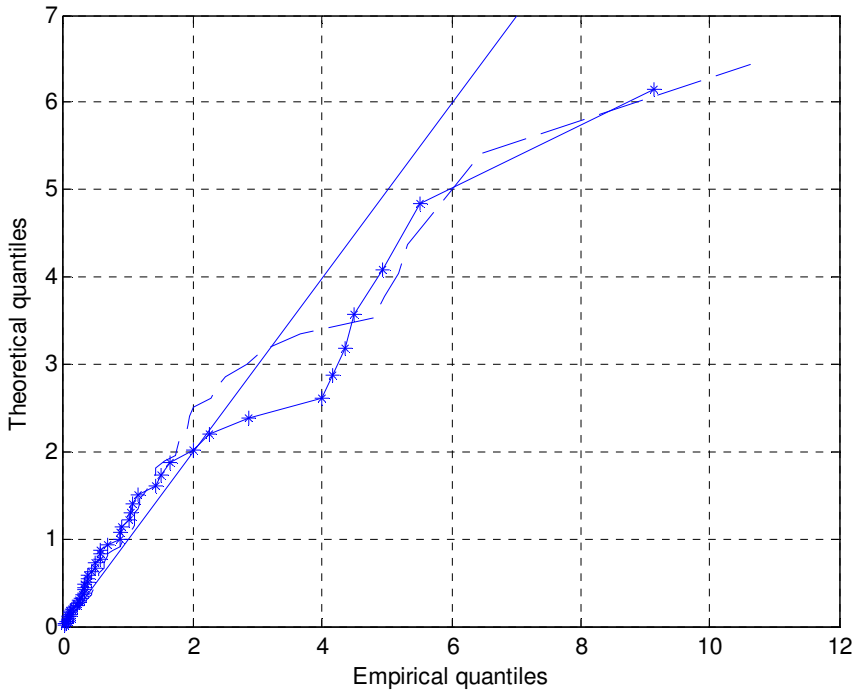
(a) Dimension 1



(b) Dimension 2



(c) Dimension 3



The line the closest to the 45° line will indicate the best fit. Although very similar, our model seems to be closer to the 45° line in figures (a) and (c). A further test for goodness of fit is the correlation coefficient r_Q between the empirical quantiles and the theoretical quantiles. The closer r_Q is to 1 the better the fit (Beirlant *et al*. 2004, p.9). The r_Q values are calculated as follows:

Table 2 Calculated r_Q values.

	Our model	**Model of Beilant et al. (2000)**
Dimension 1	0.9895	0.9811
Dimension 2	0.9746	0.9856
Dimension 3	0.9797	0.9614

The correlation coefficients are high in both cases, although it is slightly higher in with our model for dimensions 1 and 2.

8. CONCLUSION

The approach considered here for estimating the parameters of the MGBG seems to be appropriate and easier than previously reported. It seems to be a slight improvement on the estimation approach considered in Beirlant *et el*. (2000). Further research can be done in perhaps considering a thoroughly Bayesian approach for estimating all four parameters simultaneously.

APPENDIX

A.1:

The expected value of $X_i, i = 1,..., p$, where $V_{\xi_i} = \frac{1}{\xi_i}\log(1+\xi_i V_i) \sim \mathsf{GAM}(k_i, 1)$ is given by the approximation

$$E(X_i) \approx \exp(-\mu_{Y_i}) + \frac{1}{2}\sigma_{Y_i}^2 \exp(-\mu_{Y_i}).$$

(a.1)

Proof:

From (a.1) $V_i = \frac{e^{\xi_i V_{\xi_i}} - 1}{\xi_i}$, therefore $E\left[\log\left(\frac{e^{\xi_i V_{\xi_i}} - 1}{\xi_i}\right)\right] = E\left[\log(V_i)\right]$.

To obtain the $E\left[\log(V_i)\right]$ the delta method given by Rice (1995, p. 149) is used. This method involves the expansion of the Taylor series to the second order to improve the approximation. Therefore $E(\log(V_i)) \approx g(\mu_{V_{\xi_i}}) + \frac{1}{2}\sigma_{V_{\xi_i}}^2 g''(\mu_{V_{\xi_i}})$,

where

$$\mu_{V_{\xi_i}} = k_i,$$

$$\sigma_{V_{\xi_i}}^2 = k_i,$$

$$g(\mu_{V_{\xi_i}}) = \log\left(\frac{e^{\xi_i \mu_{V_{\xi_i}}} - 1}{\xi_i}\right),$$

$$g'\left(\mu_{V_{\xi_i}}\right)=\frac{\xi_i e^{\xi_i \mu_{V_{\xi_i}}}}{e^{\xi_i \mu_{V_{\xi_i}}}-1}$$

and

$$g''\left(\mu_{V_{\xi_i}}\right)=\frac{-\xi_i^2 e^{\xi_i \mu_{V_{\xi_i}}}}{\left(e^{\xi_i \mu_{V_{\xi_i}}}-1\right)^2}.$$

Therefore

$$E\left(\log\left(V_i\right)\right)\approx \log\left(\frac{e^{\xi_i k_i}-1}{\xi_i}\right)-\frac{k_i \xi_i^2 e^{\xi_i k_i}}{2\left(e^{\xi_i k_i}-1\right)^2}.$$

(a.2)

The delta method also indicates that $Var\left(\log V_i\right)\approx \sigma_{V_{\xi_i}}^2\left(g'\left(\mu_{V_{\xi_i}}\right)\right)^2$, therefore

$$Var\left(\log V_i\right)\approx k_i\left(\frac{\xi_i e^{\xi_i k_i}}{e^{\xi_i k_i}-1}\right)^2.$$

(a.3)

From (2) $Y_i=-\log\left(X_i\right)$, therefore $X_i=g\left(Y_i\right)=\exp\left(-Y_i\right)$, where

$$Y_i = -\sum_i^{\frac{1}{2}} D_\psi^{-1} \begin{pmatrix} \log\left(\frac{e^{\xi_1 V_{\xi_1}} - 1}{\xi_1}\right) - \psi(k_1) \\ . \\ . \\ . \\ \log\left(\frac{e^{\xi_p V_{\xi_p}} - 1}{\xi_p}\right) - \psi(k_p) \end{pmatrix} + \mu_i$$

$$= -\sum_i^{\frac{1}{2}} D_\psi^{-1} \begin{pmatrix} \log(V_1) - \psi(k_1) \\ . \\ . \\ . \\ \log(V_p) - \psi(k_p) \end{pmatrix} + \mu_i$$

(a.4)

and

$$D_\psi = \text{diag}\left(\sqrt{\psi'(k_1)}, \sqrt{\psi'(k_2)}, ..., \sqrt{\psi'(k_p)}\right).$$

(a.5)

Again the delta method is used to obtain an approximation for $E(X_i)$ (Rice 1995, p. 149). Thus, when following the delta method, the following is obtained:

$$E(X_i) \approx g(\mu_{Y_i}) + \frac{1}{2}\sigma_{Y_i}^2 g''(\mu_{Y_i}),$$

$$g'(Y_i) = -\exp(-Y_i) \text{ and } g''(Y_i) = \exp(-Y_i).$$

(a.6)

Therefore

$$E(X_i) \approx \exp(-\mu_{Y_i}) + \frac{1}{2}\sigma_{Y_i}^2 \exp(-\mu_{Y_i})$$

(a.7)

and

$$\mu_{Y_i} = E(Y_i) = -\Sigma_{(i)}^{\frac{1}{2}} D_\psi^{-1} E\begin{bmatrix}\log V_1\\ .\\ .\\ .\\ \log V_p\end{bmatrix} + \Sigma_{(i)}^{\frac{1}{2}} D_\psi^{-1}\begin{bmatrix}\psi(k_1)\\ .\\ .\\ .\\ \psi(k_p)\end{bmatrix} + \mu_i$$

(a.8)

where

$$E\left(\log(V_i)\right) \approx \log\left(\frac{e^{\xi_i k_i}-1}{\xi_i}\right) - \frac{k_i \xi_i^2 e^{\xi_i k_i}}{2\left(e^{\xi_i k_i}-1\right)^2}.$$

(a.9)

$\sigma_{Y_i}^2$ is the i^{th} diagonal element of

$$Cov(\underline{Y}, \underline{Y}') = \Sigma^{\frac{1}{2}} D_\psi^{-1} Cov(\log V, \log V') D_\psi^{-1} \Sigma^{\frac{1}{2}}.$$

(a.10)

Because V_{ξ_i} and $V_{\xi_{i'}}$ are independent, V_i and $V_{i'}, i \neq i'$ are also independent and therefore $Cov(\log V, \log V')$ is a diagonal matrix with $Var(\log V_i)$ on the diagonal, where $Var(\log V_i) \approx k_i \left(\frac{\xi_i e^{\xi_i k_i}}{e^{\xi_i \underline{k} i_i}-1}\right)^2$.

Therefore $E(X_i)$ is equal to

$$\exp\left[\left(\Sigma_i^{\frac{1}{2}} D_\psi^{-1}\begin{bmatrix}\log\left(\frac{e^{\xi_1 k_1}-1}{\xi_1}\right)-\frac{k_1\xi_1^2 e^{\xi_1 k_1}}{2\left(e^{\xi_1 k_1}-1\right)^2}\\ .\\ .\\ .\\ \log\left(\frac{e^{\xi_p k_p}-1}{\xi_p}\right)-\frac{k_p\xi_p^2 e^{\xi_p k_p}}{2\left(e^{\xi_p k_p}-1\right)^2}\end{bmatrix}\right)-\Sigma_i^{\frac{1}{2}} D_\psi^{-1}\begin{bmatrix}\psi_1\\ .\\ .\\ .\\ \psi_p\end{bmatrix}-\mu_i\right]+\frac{1}{2}\sigma_{Y_i}^2\exp\left[\left(\Sigma_i^{\frac{1}{2}} D_\psi^{-1}\begin{bmatrix}\log\left(\frac{e^{\xi_1 k_1}-1}{\xi_1}\right)-\frac{k_1\xi_1^2 e^{\xi_1 k_1}}{2\left(e^{\xi_1 k_1}-1\right)^2}\\ .\\ .\\ .\\ \log\left(\frac{e^{\xi_p k_p}-1}{\xi_p}\right)-\frac{k_p\xi_p^2 e^{\xi_p k_p}}{2\left(e^{\xi_p k_p}-1\right)^2}\end{bmatrix}\right)-\Sigma_i^{\frac{1}{2}} D_\psi^{-1}\begin{bmatrix}\psi_1\\ .\\ .\\ .\\ \psi_p\end{bmatrix}-\mu_i\right]$$

(a.11)

#

A.2:

The variance of $X_i, i=1,...,p$, where $V_{\xi_i}=\frac{1}{\xi_i}\log\left(1+\xi_i V_i\right)\sim\mathsf{GAM}\left(k_i,1\right)$ is given by the approximation

$$Var\left(X_i\right)=k_i\left[\exp\left(\Sigma_i^{\frac{1}{2}} D_\psi^{-1}\begin{bmatrix}\log\left(\frac{e^{\xi_1 k_1}-1}{\xi_1}\right)-\psi\left(k_1\right)\\ .\\ .\\ .\\ \log\left(\frac{e^{\xi_p k_p}-1}{\xi_p}\right)-\psi\left(k_p\right)\end{bmatrix}-\mu_i\right)\left(\left[\Sigma_i^{\frac{1}{2}} D_\psi^{-1}\left(\frac{\xi_i}{e^{\xi_i k_i}-1}\right)e^{\xi_i k_i}\right]\right)\right]^2$$

(a.12)

Proof:

From A.1 Y_i is defined as $Y_i = -\sum_i^{\frac{1}{2}} D_\psi^{-1} \begin{pmatrix} \log(V_1) - \psi(k_1) \\ . \\ . \\ . \\ \log(V_p) - \psi(k_p) \end{pmatrix} + \mu_i$.

In A.1 it was shown that $X_i = \exp(-Y_i), i = 1,..., p$, therefore

$$X_i = \exp\left(\sum_i^{\frac{1}{2}} D_\psi^{-1} \begin{pmatrix} \log\left(\frac{e^{\xi_1 V_{\xi_1}} - 1}{\xi_1} \right) - \psi(k_1) \\ . \\ . \\ . \\ \log\left(\frac{e^{\xi_p V_{\xi_p}} - 1}{\xi_p} \right) - \psi(k_p) \end{pmatrix} \right) - \mu_i$$

$$= g\left(V_{\xi_1},...,V_{\xi_p}\right).$$

(a.13)

When applying the delta method

$$X_i \approx g\left(\mu_{V_{\xi_i}}\right) + \left(V_{\xi_i} - \mu_{V_{\xi_i}}\right) \frac{\partial g\left(\mu_{V_{\xi_i}}\right)}{\partial V_{\xi_i}}.$$

(a.14)

Now it follows that:

$$Var(X_i) \approx Var\left(V_{\xi_i}\right) \left(\frac{\partial g\left(\mu_{V_{\xi_i}}\right)}{\partial V_{\xi_i}} \right)^2.$$

From A.1 we know that $\mu_{V_{\xi_i}} = k_i$ and $\sigma^2_{V_{\xi_i}} = k_i$.

Therefore $Var(X_i)$ can be approximated by (a.12).

#

A.3:

The covariance of $X_i, i = 1,..., p$ where $V_{\xi_i} = \frac{1}{\xi_i}\log(1+\xi_i V_i) \sim \text{GAM}(k_i, 1)$ is given by the approximation

$$Cov(X_t, X_{t'}) = a_t Cov\left(V_{\xi_t}, V'_{\xi_t}\right) a''_t \quad \text{(a.15)}$$

Proof:

From (a.14) $X_i \approx g\left(\mu_{V_{\xi_i}}\right) + \left(V_{\xi_i} - \mu_{V_{\xi_i}}\right)\frac{\partial g\left(\mu_{V_{\xi_i}}\right)}{\partial V_{\xi_i}}$ and therefore

$V_{\square}$

(a.16)

$$g\left(\mu_{\xi_i}\right) = exp\left[\sum_i^{\frac{1}{2}} D_\psi^{-1}\begin{bmatrix} log\left(\frac{e^{\xi_1 V_{\xi_1}} - 1}{\xi_1}\right) - \psi(k_1) \\ \vdots \\ log\left(\frac{e^{\xi_p V_{\xi_p}} - 1}{\xi_p}\right) - \psi(k_p) \end{bmatrix} - \mu_i\right]$$

where and $\mu_{V_{\xi_i}} = k_i$.

$$a_i = exp\left[\sum_i^{\frac{1}{2}} D_\psi^{-1}\begin{pmatrix} log\left(\frac{e^{\xi_1 k_1} - 1}{\xi_1}\right) - \psi(k_1) \\ \vdots \\ log\left(\frac{e^{\xi_p k_p} - 1}{\xi_p}\right) - \psi(k_p) \end{pmatrix} - \mu_i\right] \sum_i^{\frac{1}{2}} D_\psi^{-1} \frac{\xi_i e^{\xi_i k_i}}{e^{\xi_i k_i} - 1}$$

For the

$$Cov(X_t, X_{t'}) = a_t Cov\left(V_{\xi_t}, V'_{\xi_t}\right) a''_t \quad \text{(a.17)}$$

Because V_{ξ_i} and $V_{\xi_{i'}}$ are independent, $Cov\left(\underline{V_{\xi}}, \underline{V'_{\xi}}\right)$ is a diagonal matrix with $Var\left(V_{\xi_t}\right) = k_t$ in the diagonal.

#

A4:

Maximum monthly wind speed data measured at Cape Town Harbour, Cape Town Airport and Robben Island, from March 1993 to December 1998.

Harbour	Airport	Robben Island
121	45	35
49	56	37
39	41	27
41	45	35
56	60	39
56	47	37
39	49	41
49	43	37
53	47	39
53	49	43
51	47	39
51	47	39
47	43	39
60	45	37
45	45	37
89	70	53
49	54	39
43	45	43
41	45	33
43	39	31
43	45	33
47	47	37
56	51	37
49	43	35
62	41	35
51	47	64
136	51	33
70	56	45

89	51	33
58	45	41
41	45	31
54	54	45
45	47	37
54	53	41
54	49	39
49	51	31
53	47	35
49	47	31
37	45	29
51	64	41
47	49	33
97	49	41
47	53	37
56	47	41
53	51	41
51	53	37
53	49	37
58	51	39
54	45	37
47	39	41
49	47	31
58	70	45
37	43	33
72	51	41
53	45	45
45	45	37
43	49	37
56	54	41
54	51	39
54	49	39
49	45	33
47	37	29
51	47	39
43	41	29
43	47	39
45	47	35
51	47	33
49	41	37
53	43	39
54	43	39

REFERENCE

[1] Beirlant J, De Waal DJ & Teugels JL. 2000. A Multivariate Generalized Burr-Gamma Distribution. *South African Statistical Journal* **34**: 111-133.

[2] Beirlant J, De Waal DJ & Teugels JL. 2002. The generalized Burr-Gamma family of distribution with applications in extreme value analysis. *Limit Theorems in Probability and Statistics* **I**: 113-132.

[3] Beirlant J, Goedgebeur Y, Segers J & Teugels J. 2004. Statistics of Extremes Theory and Applications; Wiley & Sons: England.

[4] Conover WJ. 1980. Practical Nonparametric Statistics. Second Edition; Weily & Sons: United States of America.

[5] Meng XL. 1997. The EM algorithm and medical studies: a historical link. *Statistical Methods in Medical Research* **6:** 3-23.

[6] Rice JA. 1995. Mathematical Statistics and Data Analysis. Second edition; Duxbury Press: California.

[7] Salisbury JI & Wimbush M. 2002. Using modern time series analysis techniques to predict ENSO events from the SOI time series. *Nonlinear Processes in Geophysics* **9**: 341-345.

Semi-Parametric Inference for Measures of Inequality (Preliminary Report)

Tertius de Wet[1]
tdewet@sun.ac.za
Stellenbosch University

Abstract

In this paper we discuss the semi-parametric estimation of measures of inequality, in particular, the Gini measure. This estimation procedure is specifically applicable in the case of heavy tailed distributions. Such distributions often occur in real data sets e.g. in income data which usually have a heavy right tail. The estimation is illustrated by application to the South African Income and Expenditure data of 2005.

1. Introduction.

Measures of inequality, also called measures of concentration or diversity, are very popular in economics and especially in measuring the inequality in income or wealth within a population and between populations. However, they have applications in many branches of science, e.g. in ecology (Magurran, 1991), linguistics (Herdan, 1966), sociology (Allison, 1978), demography (White, 1986) and information science (Rousseau, 1993).

Over the years a large number of measures have been proposed, with the Gini index perhaps the most well-known one. Having reliable inequality measures available is an important first step. A next step is to estimate the values of these measures using samples from the appropriate populations and, in particular, to estimate the variability of these estimators and more generally, to obtain confidence intervals for the measures. Since inequality is inherently dependent on the tails of a population, estimators of inequality are typically sensitive to data from these tails. We note in this regard that all the well-known inequality measures have unbounded influence functions. It is well-known that income distributions often exhibit a long tail to the right, making estimators of inequality particularly sensitive to large values. It is thus important to study the behaviour of estimators based on data from heavy tailed distributions. Many of the traditional estimators are sensitive to such extreme data points (see e.g. Cowell and Flachaire, 2007) and remedial action needs to be taken. This remedial action can be either a trimming of the extreme data or a modification of the estimator to make it more robust to extreme observations.

Cowell and Flachaire (2007) (see also Cowell and Victoria-Feser, 2008) have proposed a so-called semi-parametric approach to modify estimators under heavy tailed distributions. This method estimates the left part of the distribution, where the bulk of the distribution resides, using the usual nonparametric empirical distribution function and the right (upper) part of the distribution using a parametric extreme value distribution, e.g. a Pareto distribution. The resulting estimator is therefore partly nonparametric and partly parametric, hence semi-parametric. Results from extreme value theory are then used to obtain a more reliable distribution estimator. This new estimator of the distribution forms the basis for more robust estimators of the measures of inequality. Since these new estimators are expected to have a limiting

[1] This paper is based on joint work with Tchilabalo Kpanzou and Ariane Neethling, both from Stellenbosch University.

normal distribution, approximate confidence intervals can be obtained using standard normal theory.
In this (preliminary) report we build on these ideas and focus, in particular, on the Gini index for illustrative purposes. It is shown how results from extreme value theory can be used to find an appropriate parametric distribution to use, also leading to a choice of the number of observations to use in fitting this parametric distribution. The results are applied to a well-known South African income data set, the IES 2005.
The layout of the paper is as follows. In the next section a number of measures of inequality are discussed, in Section 3 the approach to semi-parametric estimation of the Gini index is discussed and in Section 4 the results are applied to the IES 2005 data set. A number of further aspects being investigated are discussed in the closing section.

2. Measures of Inequality.

A large number of measures of inequality have been proposed in the literature. We discuss the two most important ones, viz. the generalized entropy (class of) measures, which includes the Theil and mean logarithmic deviation, and the Gini measure. In addition, we discuss a more recent proposal, the quintile share ratio. Our focus for this paper will be on the Gini inequality measure.
Let Y denote the random variable of interest and suppose Y is positive valued and has a distribution function F.

2.1 Generalized Entropy Measures

The generalized entropy (GE) measure is defined by (see e.g. Cowell and Flachaire, 2007)

$$I_E^\alpha = \left[\alpha(\alpha-1)\right]^{-1}\int_0^\infty\left[\left(\tfrac{y}{\mu}\right)^\alpha - 1\right]dF(y) = \left[\alpha(\alpha-1)\right]^{-1}\left(\tfrac{\upsilon}{\mu^\alpha}-1\right),$$

where

$$\mu = \int_0^\infty y dF(y),\ \upsilon = \int_0^\infty y^\alpha dF(y) \text{ and } \alpha \in R.$$

It follows quite easily that the influence function of I_E^α is given by

$$IF\left(z; I_E^\alpha\right) = \left[z^\alpha - \upsilon\right] - \tfrac{\upsilon}{(\alpha-1)\mu^{\alpha+1}}\left[z-\mu\right].$$

Clearly the influence function is unbounded when $z \to \infty$.
Two special cases of the GE measure are the Theil measure, when $\alpha = 1$, i.e.

$$I_E^1 = \tfrac{\upsilon}{\mu} - \log\mu,$$

where now $\upsilon = \int_0^\infty y \log y dF(y)$,

and the mean logarithmic deviation, when $\alpha = 0$, i.e.

$$I_E^0 = \int_0^\infty \log\left(\tfrac{y}{\mu}\right)dF(y) = \log\mu - \upsilon,$$

where now $\upsilon = \int_0^\infty \log y dF(y)$.

2.2 Gini Index

The Gini index is the most widely used measure of inequality. There are a number of different ways of defining it, the most well-known the following.

Let
$Q(q;F) \equiv \inf\{y : F(y) \geq q\}$
denote the quantile function of F,
$C(q;F) \equiv \int_0^q Q(u;F)\,du = \int_0^{Q(q;F)} y dF(y)$,

the cumulative quantile function and
$\mu = \int_0^\infty y dF(y)$,

as before.
The Gini index is then given by

$I_G \equiv 1 - 2\mu^{-1}\int_0^1 C(u;F)\,du \equiv 1 - 2R(F)$,

with
$R(F) = \mu^{-1}\int_0^1 C(u;F)\,du$.

Remark. The Gini index is closely related to the Lorenz curve. The latter is defined as
$L(q;F) \equiv \frac{C(q;F)}{\mu}$.
Thus
$I_G = 1 - 2\int_0^1 L(u;F)\,du$,
i.e. one minus twice the area under the Lorenz curve. Furthermore

$I_G = 2\left(\int_0^1 u du - \int_0^1 L(u;F)\,du\right) = 2\left(\int_0^1 (u - L(u;F))\,du\right)$.

This shows that the Gini index is twice the area between the 45 degree line and the Lorenz curve.

The influence function of the Gini index is easily seen to be
$IF(z;I_G) = 2\left[R(F) - C(F(z);F) + \frac{z}{\mu}(R(F) - (1 - F(z)))\right]$,
which tends to infinity at the rate of z when $z \to \infty$.

A number of alternative definitions of the Gini index are as follows.
(i) $I_G = \frac{1}{2\mu} E|Y_1 - Y_2|$,
where Y_1 and Y_2 are independent random variables each having distribution F.
(ii) The expression in (i) can also be written as

$$I_G = \tfrac{1}{2\mu}\int_0^1\int_0^1 |Q(u;F)-Q(v;F)|\,dudv\,.$$

(iii) $I_G = 1-\tfrac{1}{\mu}\int_0^\infty \left(1-F(y)\right)^2 dy\,.$

Note that this can also be written as

$$I_G = 1-\tfrac{1}{\mu}\int_0^\infty \left(1-F(y)\right)^2 dy = \mu^{-1}\int_0^\infty F(y)\left(1-F(y)\right)dy\,.$$

Remark. Note that the Gini index lies between zero and one, with zero indicating perfect equality and one indicating perfect inequality.

2.3 Quintile Share Ratio Measure

This measure is defined as the ratio of the 0.8^{th} quantile to the 0.2^{th} quantile, i.e. $QSR = Q(0.8;F)/Q(0.2;F)$.

Remark. This measure forms part of the so-called Laeken indicators, the European indicators on poverty and social exclusion. We only mention this measure and will not discuss it further.

3. Semi-parametric estimation for the Gini index.

Consider now the Gini index as given in (iii) in the previous section:

$$I_G = \mu^{-1}\int_0^\infty F(y)\left(1-F(y)\right)dy\,.$$

Given a sample $Y_1,...,Y_n$ of size n on Y, the usual nonparametric estimator for I_G is given by

$$\hat{I}_G^{NP} = \hat{\mu}_n^{-1}\int_0^\infty F_n(y)\left(1-F_n(y)\right)dy\,,$$

where

$$F_n(y) = n^{-1}\sum_{i=1}^n I(Y_i \le y)$$

is the empirical distribution function and

$$\hat{\mu}_n = n^{-1}\sum_{i=1}^n Y_i$$

is the sample average.

In many cases the distribution function F has a heavy right tail, with relatively fewer observations in the tail part of the distribution. In such a situation it may not be wise to use the empirical distribution function to estimate F over its whole range. A better choice might be to use F_n to estimate F over the bulk of the data, but to use a parametric estimator in the tail part of F. This was the approach suggested by Cowell and Flachaire (2007) (see also Cowell and Victoria-Feser, 2008). Since it is assumed that F has a heavy right tail, extreme value theory can be used. In this section such an approach will be followed.

Now, define a semi-parametric distribution function by

$$\widetilde{F}(y) = \begin{cases} F(y), & y \le y_0 \\ F(y)+\left(1-F(y)\right)F_\theta(y), & y > y_0 \end{cases}$$

for a given y_0, where F_θ is derived from some parametric distribution (e.g. Pareto or generalized Pareto). The distribution function F can be thought of as the unknown underlying distribution function that will be estimated (in the bulk of the data) by the empirical distribution function.

Taking y_0 as an upper quantile of F, say $y_0 = Q(1-\alpha;F)$, leads to

$$\widetilde{F}(y) = \begin{cases} F(y), & y \le Q(1-\alpha;F) \\ 1-\alpha\left(1-F_\theta(y)\right), & y > Q(1-\alpha;F). \end{cases}$$

The semi-parametric Gini index is given by

$$I_{SPG} = \widetilde{\mu}^{-1}\int_0^\infty \widetilde{F}(y)\left(1-\widetilde{F}(y)\right)dy.$$

We now give a simplified expression for I_{SPG} that will be used in deriving an estimator for it.

First, note that

$$\widetilde{\mu} = \int_0^\infty \left(1-\widetilde{F}(y)\right)dy$$

$$= \int_0^{Q(1-\alpha;F)} \left(1-F(y)\right)dy + \alpha\int_{Q(1-\alpha;F)}^\infty \left(1-F_\theta(y)\right)dy$$

and

$$I_{SPG} = \widetilde{\mu}^{-1}\int_0^\infty \widetilde{F}(y)\left(1-\widetilde{F}(y)\right)dy$$

$$= \widetilde{\mu}^{-1}\left[\int_0^{Q(1-\alpha;F)} F(y)\left(1-F(y)\right)dy + \alpha\int_{Q(1-\alpha;F)}^\infty \left[1-\alpha\left(1-F_\theta(y)\right)\right]\left(1-F_\theta(y)\right)dy\right].$$

On the interval $\left(0, Q(1-\alpha;F)\right)$ we estimate F by the empirical distribution function F_n.

Now, denote the order statistics of the sample by $Y_{1,n} < Y_{2,n} < \ldots < Y_{n,n}$ and let $\alpha = \frac{k}{n}$ for some appropriately chosen integer k. Then $Q(1-\alpha;F)$ can be estimated by $Y_{n-k,n}$.

On the interval $\left(Q(1-\alpha;F), \infty\right)$ we want to use a parametric distribution obtained from extreme value theory. This approach is now briefly described.

Write GPD for the generalized Pareto distribution and GEV for the generalized extreme value distribution. It is then well known (see e.g. Beirlant et al., 2004) that for a threshold u

$D(Y-u \mid Y>u) \doteq GPD$, for large u,

whenever Y lies in the domain of attraction of the GEV.

Therefore, if G denotes the GPD, we can write

$$P[Y-u > y \mid Y > u] = P[Y-u > y] / P[Y > u]$$

$$= \overline{F}(y+u) / \overline{F}(u)$$

$$\approx \overline{G}(y),$$

where $\overline{F}(y) = 1-F(y)$ and $\overline{G}(y) = 1-G(y)$.

It follows that
$\overline{F}(y+u) \approx \overline{F}(u)\overline{G}(y)$,
which becomes (by letting $y+u \to y$)
$\overline{F}(y) \approx \overline{F}(u)\overline{G}(y-u)$ and
$F(y) \approx 1-\overline{F}(u)\overline{G}(y-u)$.
The right hand side of this is then the $F_{\underline{\theta}}(y)$ in the semi parametric form above.

Remark. Note that the GPD has two unknown parameters, say γ and σ. The form of G is then
$G(x) = 1-\left[1+\gamma\frac{x}{\sigma}\right]^{-1/\gamma}$.
The parameters γ and σ can be estimated by maximum likelihood. See Beirlant et al. (2004).
We now give an estimator for I_{SPG} by estimating the elements in its expression given above.
Note that, as mentioned above, if we take $\alpha = \frac{k}{n}$ then $Q(1-\alpha; F)$ can be estimated by $Y_{n-k,n}$. Therefore, for $y \le Y_{n-k,n}$ we estimate $\widetilde{F}$ by
$\widetilde{F_n}(y) = F_n(y)$,
the empirical distribution function.
Consider now $y > Y_{n-k,n}$ and let $\underline{\theta} \equiv (\gamma, \sigma)'$, the GPD parameter and write $\hat{\underline{\theta}}$ for the corresponding maximum likelihood estimator.
Writing $G \equiv G_{\underline{\theta}}$ to indicate the dependence on the unknown parameter, we estimate it by $G_{\hat{\underline{\theta}}}$. Using the above approximation for $F(y)$ for a large threshold, for $y > Y_{n-k,n}$ we estimate $\widetilde{F}$ by
$\widetilde{F_n}(y) = 1-\frac{k}{n}\left[1-\left(1-\frac{k}{n}\overline{G_{\hat{\underline{\theta}}}}\left(y-Y_{n-k,n}\right)\right)\right]$.
Here, with the threshold u taken as $Y_{n-k,n}$, the estimator for $\overline{F}(u)$ is taken as $1-F_n\left(Y_{n-k,n}\right) = \frac{k}{n}$.
Putting this together, gives the final estimator as

$$\widetilde{F_n}(y) = \begin{cases} F_n(y) & for\ y \le Y_{n-k,n} \\ 1-\frac{k}{n}\left[1-\left(1-\frac{k}{n}\overline{G_{\hat{\underline{\theta}}}}\left(y-Y_{n-k,n}\right)\right)\right] & for\ y > Y_{n-k,n}. \end{cases}$$

Substituting this in I_{SPG}, gives our estimator $\hat{I}_{SPG}$. The following result gives a convenient computational formula for this estimator.

Theorem. The estimator $\hat{I}_{SPG}$ can be written as

$$\hat{I}_{SPG} = \left(n\hat{\mu}\right)^{-1}\sum_{i=1}^{n-k} i\left(1-\tfrac{i}{n}\right)\left(Y_{i,n}-Y_{i-1,n}\right) + k^2\hat{\sigma}\left[2n^2-k^2-\hat{\gamma}\left(n^2-k^2\right)\right] / n^4\hat{\mu}\left(1-\hat{\gamma}\right)\left(2-\hat{\gamma}\right),$$

where $\hat{\gamma}$ and $\hat{\sigma}$ are the maximum likelihood estimators and $\hat{\mu}$ can be calculated as

$$\hat{\mu} = \sum_{i=1}^{n-k} \left(1 - \tfrac{i}{n}\right)\left(Y_{i,n} - Y_{i-1,n}\right) + k^2 \hat{\sigma} / n^2 \left(1 - \hat{\gamma}\right), \quad Y_{0,n} = 0, \hat{\gamma} < 1..$$

Remark. As is usual in extreme value theory, the choice of the threshold u to work with, or equivalently the choice of the number of exceedances k, is of critical importance. In our application we will apply a goodness-of-fit test to test for the GPD and choose k as the value that gives a reasonable p value. The goodness-of-fit test used will be based on the work of Choulakian and Stephens (2001) and Villasenor-Alva and Gonzalez-Estrada (2009).

4. Results for IES 2005.

We apply the proposed method to the variable EQ_INC (Equalized Income) of the South African Income and Expenditure Survey data (IES2005), consisting of 20 986 observations. The data ranges from 0 to 2 910 826.00, with a heavy right tail as shown by the histogram below.

Applying the GPD test (as discussed in Choulakian and Stephens, 2001 and Villasenor-Alva and Gonzalez-Estrada, 2009) to the exceedances, we found a p-value of $\hat{p} = 0.78$ for $k = 630$ (3% of the data). The estimates of the shape and scale parameters for the GPD are respectively given by $\hat{\gamma} = 0.6402$ and $\hat{\sigma} = 45053.7500$. The corresponding semi-parametric Gini index is estimated by $\hat{I}_{SPG} = 0.6154$. The nonparametric estimate of the Gini index is given by $\hat{I}_G = 0.6703$.

5. Further work.

This is a preliminary report on a research project that has only recently started. Clearly there are still many outstanding issues to explore. Below some of these are mentioned that are currently being investigated.

- If instead of the exceedances $Y - u$, relative exceedances Y / u are considered, then it is well known that
  $D\left(Y / u \mid Y > u\right) \approx Pareto,$ for large u.
  One can thus carry out the program above for the Pareto distribution rather than the GPD. Furthermore, a second order approximation exists for such exceedances, viz. approximation by the so-called perturbed Pareto distribution (PPD). This should give a better fit in the tail of the distribution.
- The influence functions of the measures of inequality considered are clearly sensitive to outlying values. What is the effect of outliers on the semi-parametric estimators? How should this effect be countered, e.g. through appropriate trimming?
- The above was illustrated only for the Gini measure. Similar results are being derived for other measures of inequality.
- Estimating a measure of inequality is only a first step. The next step would be to obtain confidence intervals. This is being carried out in different ways.
  - Using a re-sampling method such as the bootstrap.

- Using asymptotic theory. Since these estimators are asymptotically normally distributed this is fairly straightforward. However, results may not be accurate for "small" samples.
- In order to improve small sample accuracy of the confidence intervals, use could be made of transformations. Some work in this direction has recently been done by Schluter and van Garderen (2009).

- Survey data is often based on complex sampling schemes. How can the semi-parametric approach be modified to accommodate such data?

6. References.

Allison P.D. (1978). Measures of inequality. *American Sociological Review* **43**, 865 – 880.

Beirlant J., Goegebeur Y., Segers J. and Teugels J. (2004). *Statistics of Extremes: Theory and Applications.* Wiley, New York.

Choulakian V. and Stephens M.A. (2001). Goodness-of-Fit Tests for the Genaralized Pareto Distribution. *Technometrics* **43**, 478 – 484.

Cowell F.A. and Flachaire E. (2007). Income Distribution and Inequality Measurement: The Problem of Extreme Values. *Journal of Econometrics* **141**, 1044 – 1072.

Cowell F.A. and Victoria-Feser M-P. (2008). Modelling Lorenz Curves: Robust and Semi-Parametric Issues. In Duangkomen Chotikapanich (editor) *Modelling Income Distributions and Lorenz Curves*, 241 – 253. Springer.

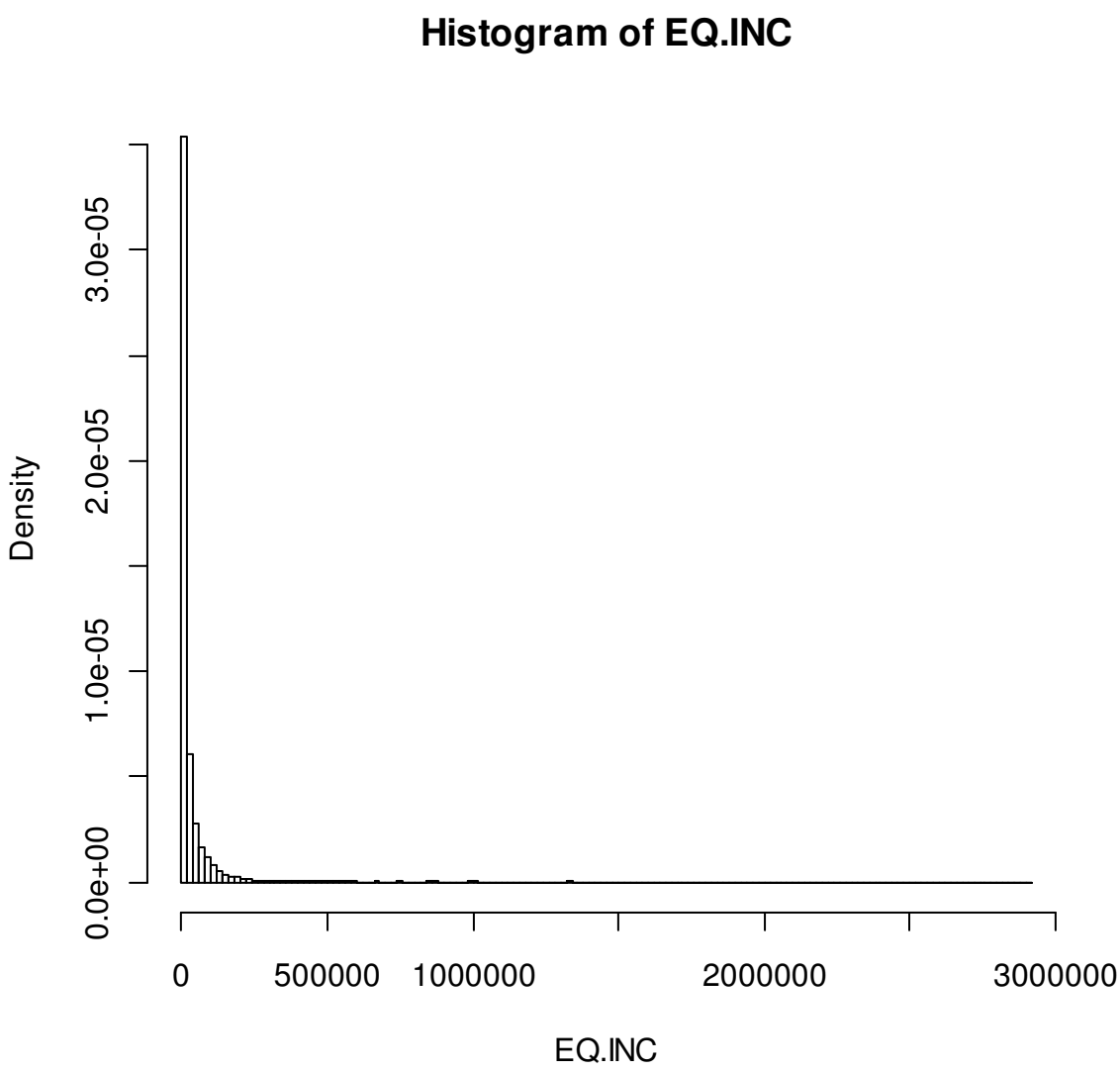
Herdan G. (1966). *The Advanced Theory of Language as Choice and Chance*. Springer.

Rousseau R. (1993). Measuring Concentration: Sampling design issues, as illustrated by the case of perfectly stratified samples. *Scientometrics* **28**, 3 – 14.

Schluter C. and van Garderen K.J. (2009). Edgeworth Expansions and Normalizing Transforms for Inequality Measures. *Journal of Econometrics* **150**, 16 – 29.

Villasenor-Alva J.A. and Gonzalez-Estrada E. (2009). A Bootstrap Goodness of fit Test for the Generalized Pareto Distribution. *Computational Statistics and Data Analysis* **53**, 3835 – 3841.

White M.J. (1986). Segregation and Diversity Measures in Population Distribution. *Population Index* **52**, 198 – 221.


Histogram of EQ.INC
Density
0.0e+00
1.0e-05
2.0e-05
3.0e-05
0
500000
1000000
2000000
3000000
EQ.INC

AN APPLICATION OF SEQUENTIAL REGRESSION MULTIPLE IMPUTATION ON PANEL DATA

MICHAEL JOHAN VON MALTITZ
ABRAHAM JOHANNES VAN DER MERWE

1. INTRODUCTION

One of the major issues associated with large surveys is that of non-response or lost data. Moreover, the main problematic issue regarding missing data is that most data analysis procedures are not designed to handle them, leading to analyses that conclude invalid and inefficient inferences about a population (Schafer and Graham, 2002). Many economic analyses use either complete-case analysis or a simple but inaccurate method of imputing missing data. In most cases, the missing data is multivariate, meaning that missing values appear in several observed variables. In essence, we have that a complete data matrix Y_{com} is made up of an observed part, Y_{obs}, and a missing part, Y_{mis}. The data that is missing has a particular distribution of positions. These positions are indicated by a matrix R of zeroes and ones that has the same dimensions as the complete data matrix Y_{com}. In R *t*here are ones in the positions of the missing data entries of Y_{com}, and zeroes elsewhere. The distribution of R, referred to as the distribution of 'missingness', is $\mathrm{P}(\mathrm{R}|\mathrm{Y}_{com}, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a vector of unknown parameters.

There are three mechanisms by which data is said to be missing – 'missing at random' (MAR), 'missing completely at random' (MCAR), or 'missing not at random' (MNAR). In the MAR mechanism, the distribution of positions of the missing data entries is assumed to be independent of the missing data in the analysis, or $\mathrm{P}(\mathrm{R}|\mathrm{Y}_{com}, \boldsymbol{\xi}) = \mathrm{P}(\mathrm{R}|\mathrm{Y}_{obs}, \boldsymbol{\xi})$. In the case of MCAR, a special version of the MAR mechanism, the positions of the missing data entries are assumed to be independent of all of the variables in the analysis, i.e. $\mathrm{P}(\mathrm{R}|\mathrm{Y}_{com}, \boldsymbol{\xi}) = \mathrm{P}(\mathrm{R}|\boldsymbol{\xi})$. In the last case, the MNAR missing data mechanism, the positions of the missing data entries are assumed to be at least dependent on data that is missing from the dataset, or, more basically, the distribution of missingness is not MAR. This means that for MNAR, $\mathrm{P}(\mathrm{R}|\mathrm{Y}_{com}, \boldsymbol{\xi}) \neq \mathrm{P}(\mathrm{R}|\mathrm{Y}_{obs}, \boldsymbol{\xi})$. If the mechanism behind the missing data is MAR, then the mechanism is said to be 'ignorable' (Little and Rubin, 1987).

In complete-case analysis, only cases containing values for each of the variables in question are retained in the data analysis procedures. This can raise the problem of serious bias in the analysis (Little and Rubin, 1987). One must note, however, that these possible biases may not always exist in complete-case analysis, but rather that the extent of bias will depend on the mechanism by which data is deemed to be missing. Particularly, if the data is MCAR, then there will be no bias in complete-case analysis of multivariate data with missing entries (Schafer, 2003). To overcome the possible biases in complete-case analysis, many methods of dealing with incomplete data and imputing missing values have been suggested. The non-imputation methods of handling incomplete data include available-case analysis, weighting procedures or model-based procedures.

Alternatively, if complete-case analysis methods are to be used on a dataset that is originally incomplete, data can be 'filled in' by several imputation procedures, including substitution, cold-deck imputation, unconditional and conditional mean substitution, imputation from unconditional distributions or hot-deck imputation, and imputation from conditional distributions. All of these imputation procedures are single imputation methods, imputing only one value for each missing datum.

Multiple imputation, also proposed by Rubin (1987), is viewed as a flexible alternative to likelihood methods for a range of incomplete data problems (Schafer and Graham, 2002). The primary advantage of multiple

imputation is the inflation of uncertainty in the analysis estimates. In essence multiple imputation covers a class of methods that impute several plausible values for a single missing data entry. Once the missing values have been imputed, several completed datasets are left to be analysed by complete-case methods. A simple set of rules is then used to combine the estimates from the separate analyses of the several datasets, and the uncertainty of these estimates is then formed from the sample variation as well as variation in the imputed values themselves. Although the number of datasets that should be completed is often debated, a small number of completed datasets, say, between 10 and 20, often suffices in order to obtain precise estimates.

Multiple imputation can be easily linked to Bayesian statistics, as the imputed values for a single missing data entry can be draws from the predictive posterior distribution for the non-missing data. One relatively new method of imputation is that of Raghunathan *et al.* (2001), namely sequential regression multiple imputation, or SRMI. This method extends and refines imputation from conditional distributions into a multiple imputation context. The process will be detailed in the methodology section.

2. METHODOLOGY

2.1 Hypothesis

The hypothesis of interest is that there are no changes in the parameters of a regression analysis once imputation has been performed using the SRMI process. More accurately, we want to determine whether or not social networks have influenced welfare in South Africa. The research problem is quite simply linked to this hypothesis. If there are changes in the parameters of interest from before imputation to after imputation, then the researcher has to decide which set of parameters, if any, will be used to aid policy decision. The problem to be solved is the adequacy of whichever estimates are decided on.

2.2 Data

This research uses the combined datasets of the Project for Statistics on Living Standards and Development (PSLSD), undertaken in 1993, and the follow up KwaZulu-Natal Income Dynamics Study (KIDS) surveys undertaken in 1998 and 2004.[1] These two studies form a three-period panel data set of 1412 households in 1993, 1075 households in 1998, and 1428 households in 2004. Once data constraints have been taken into account, the household models include 1158 African and Indian households. These figures are calculated before cases are dropped in any complete-case analysis on the data. Additionally, since the 1993 survey was designed to be self-weighted, it may only be assumed that the 1998 sample of households is representative of the 1993 population, as is the sample in 2004. The data actually spans 3928 observations in 1906 households when the survey waves are combined in a single dataset. If a complete-case random effects regression is performed on this data, 1600 cases are dropped before the analysis is completed. This is a large number relative to the number of households available for analysis, meaning that the complete-case procedures may produce severely biased results if the data is not MCAR.

2.3 Multiple Imputation through Sequential Regression

Raghunathan *et al.* (2001: 85) summarize the SRMI process for a single cross-sectional dataset:

[1] The KwaZulu-Natal Income Dynamics Study (KIDS) was a collaborative project of the International Food Policy Research Institute, the University of Natal-Durban, the University of Wisconsin-Madison, and the Southern Africa Labour and Development Research Unit at the University of Cape Town. The 2004 data used in the analysis was generously provided by the KIDS research team before the public release of the data. This means that the data had yet to be cleaned. Even though extensive cleaning of the data was required, in particular for individual's ages, education levels, household heads, and even gender, the cleaning was approached with much care, so as to preserve the validity of the data.

> "This approach specifies an explicit model for variables with missing values, conditional on the fully observed variables and some unknown parameters, and a model for the missing data mechanism, which does not need to be specified under an ignorable missing data mechanism (Rubin, 1976). This explicit model then generates a posterior predictive distribution of the missing values conditional on the observed values. The imputations are draws from this posterior predictive distribution."

Firstly, the dataset's incomplete variables are sorted from the variable with the least missing entries to the variable with the most missing values. Let the variable with the least missing be $\boldsymbol{y}_1$, the variable with the next fewest missing be $\boldsymbol{y}_2$, *etc*., until $\boldsymbol{y}_k$. Let X be that part of the dataset that is originally complete. Finally, let $\boldsymbol{\xi}_j$ be a vector of the unknown regression and dispersion parameters in the conditional model for $\boldsymbol{y}_j$. The sorting of the dataset follows as an extension to the fact that in model-based imputations the joint conditional density of $\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_k$ given X can be factored as

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_k|\mathrm{X}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_k) \\ = f_1(\boldsymbol{y}_1|\mathrm{X}, \boldsymbol{\xi}_1) f_2(\boldsymbol{y}_2|\mathrm{X}, \boldsymbol{y}_1, \boldsymbol{\xi}_2) \dots f_k(\boldsymbol{y}_k|\mathrm{X}, \boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_{k-1}, \boldsymbol{\xi}_k) \quad (1.1)$$

Each conditional density is modelled by an appropriate regression model with unknown parameters, $\boldsymbol{\xi}_j$.

Secondly, the first round of imputations begins, and the variable with the least missing data entries (apart from the complete variables) is selected. This variable is regressed on the complete data according to a regression model that is assumed to fit the distribution of the variable, as mentioned above. The model first processed is illustrated in Equation (2.1) by f_1. The regression is Bayesian by nature, but utilizes a diffuse or non-informative prior. If $\xi = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_k)$, then the prior for each model is $\pi(\xi) \propto 1$. A set of regression parameters is then drawn from the regression model and a single draw from the predictive posterior of the model (the predictive distribution of the missing values given the observed values) is made for every missing data entry in that variable. These draws are the imputed values for that variable.

Thirdly, the SRMI process then selects the variable with the next fewest missing values, and the procedures in the second step are repeated. A new regression model, illustrated by f_2 in Equation (2.1), is chosen according to the assumed distribution of $\boldsymbol{y}_2$, the variable now being regressed. This new variable is regressed on the complete data *and* the newly completed variable from the previous step (i.e. the variable with the least missing values, all of which have now been imputed with a single imputation). Again a set of regression parameters is drawn from the new regression model and a single draw from the predictive posterior of this model is made for every missing data entry in the variable. This step is repeated until all of the variables in the dataset are 'filled in' by appropriate regression predictions. By the nature of this process, the terms 'sequential regression imputation' are justified.

Fourthly, once an entire dataset has been completed or updated with imputed values for the original missing data entries, this completed dataset is subjected to an update round, round two, starting essentially at the second step above. Thus, the iterative process involved in SRMI is brought to light. The process involved in the updating rounds differs slightly to that of steps two and three above.

The first difference depends on the pattern of the missing data. For a monotone pattern of missing data, if a datum for an observation is missing in variable $\boldsymbol{y}_j$, then the data for that observation will be missing in variables $\boldsymbol{y}_{j+1}, \boldsymbol{y}_{j+2}, \dots, \boldsymbol{y}_k$. When this pattern occurs the imputations in the first round are approximate draws from the predictive distribution of the missing values given the observed values. Draws in subsequent rounds can be improved upon using the SIR (sampling, importance-weighting, resampling) or another

rejection algorithm (Raghunathan *et al*., 2001). When the pattern of missing data is not monotone, a Gibbs sampling algorithm must be developed to or improve upon the previous round's estimates. Raghunathan *et al*. (2001) suggest that the missing values in $\boldsymbol{y}_j$ at round $(w+1)$ need to be drawn from the conditional density:

$$f_j^*\left(\boldsymbol{y}_j \middle| \boldsymbol{\xi}_1^{(w+1)}, \boldsymbol{y}_1^{(w+1)}, \dots \boldsymbol{\xi}_{j-1}^{(w+1)}, \boldsymbol{\xi}_{j+1}^{(w)}, \dots, \boldsymbol{\xi}_k^{(w)}, \boldsymbol{y}_k^{(w)}, \mathrm{X}\right), \tag{1.2}$$

where $\boldsymbol{y}_i^{(w)}$ is the imputed or observed value for variable $\boldsymbol{y}_i$ at round w.

Equation (2.2) is computed based on the joint distribution in Equation (2.1), This draw process would be extremely difficult to complete, since the density in Equation (2.2) is difficult to compute in most practical situations without restrictions (Raghunathan *et al*., 2001). However, Raghunathan *et al*. (2001) propose that instead, the draw in round $(w+1)$ for $\boldsymbol{y}_j$ is taken from the predictive distribution corresponding to the conditional density:

$$g_j\left(\boldsymbol{y}_j \middle| \boldsymbol{y}_1^{(w+1)}, \boldsymbol{y}_2^{(w+1)}, \dots, \boldsymbol{y}_{j-1}^{(w+1)}, \boldsymbol{y}_{j+1}^{(w)}, \dots, \boldsymbol{y}_k^{(w)}, \mathrm{X}, \boldsymbol{\phi}\right) \tag{1.3}$$

where $\boldsymbol{\phi}$ is a vector of the unknown regression parameters with diffuse prior.

In other words, in imputation rounds after the first the values that were originally missing in each variable are now predicted from regression models regressing those variables on *all* of the other variables in the dataset, meaning that the variables with values imputed from the first round are used as regressors in the second round in addition to the newly updated variables from the current round. This process can be viewed as an approximation to the Gibbs sampling procedure in Equation (2.2). In some particular cases this approximation is equivalent to drawing values from a posterior predictive distribution under a fully parametric model. For example, if all of the variables are continuous and Normally distributed with constant variance, then the algorithm governing Equation (2.3) converges to a joint predictive distribution under a multivariate Normal distribution with an improper prior for the mean and covariance matrix (Raghunathan *et al*., 2001).

This fourth step is then repeated as many times as the researcher deems fit (usually to a point where the inferences made on the data during subsequent rounds converge). The extent to which the process is repeated will be expanded upon in the following section.

2.4 Inference on the Completed Datasets

Once multiple datasets have been imputed from the same starting point, inferences on the datasets can be combined using a simple set of rules known as *Rubin's* Rules, as given by Rubin (1987), and explained below.

Suppose that Q is a scalar population quantity to be estimated from the sample data taken in a survey, and that an estimate $\hat{Q}$ and standard error $\sqrt{U}$ could be easily calculated if Y_{mis} were available. In multiple imputation, Y_{mis} is replaced by $m > 1$ simulated versions, $\mathrm{Y}_{mis}^{(1)}, \mathrm{Y}_{mis}^{(2)}, \dots \mathrm{Y}_{mis}^{(m)}$, leading to m estimates and their respective standard errors, $\left(\hat{Q}_j, \sqrt{U_j}\right), j = 1, \dots, m$. An overall estimate for Q is:

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m} \hat{Q}_j \tag{1.4}$$

with a standard error of $\sqrt{T}$, where

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B, \tag{1.5}$$

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m} U_j ,$$

and

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{Q}_j - \bar{Q}\right)^2 .$$

If we have it that $\frac{\hat{Q}-Q}{\sqrt{\bar{U}}}$ is approximately $N(0,1)$ with complete data, as is assumed to be the case in many regression contexts, then $\frac{\hat{Q}-Q}{\sqrt{\bar{U}}} \sim t_v$, in the imputed data case, where:

$$v = (m-1)\left(1 + \frac{1}{r}\right)^2 ,$$

and

$$r = \left(1 + \frac{1}{m}\right)\frac{B}{\bar{U}}.$$

The latter, r, is the relative increase in variance due to nonresponse (Schafer and Graham, 2002).

The degrees of freedom vary from $(m-1)$ to ∞ according to the rate of missing information in the dataset. This rate of missing information is given by:

$$\gamma = \frac{\left(r + \frac{2}{v+3}\right)}{r+1}, \qquad (1.6)$$

where r is as above.

Schafer and Graham (2002) also note that when the degrees of freedom is large (or alternatively when the variation in the estimates between imputations is small compared to the overall variation), then there is not much that can be gained from increasing m, the number of imputed datasets.

In order to determine the actual number of imputed datasets that should be created, Rubin (1987) also provides a measure of efficiency, measured in standard errors, and based on the rate of missing information, γ. It is given by:

$$\lambda = \left(1 + \frac{\gamma}{m}\right)^{-\frac{1}{2}} \qquad (1.7)$$

This measure essentially compares the size of the standard error after m imputations with the size of the standard error after an infinite number of imputations. This measure is used as a basis for determining whether or not the number of imputations made in Section 3's application of SRMI on real data is adequate.

2.5 Imputing from Generalised Linear Models (GLMs)

From Equation (2.3) it is evident that a particular regression model needs to be utilized, according to the assumed distribution of the variable in question, in order to obtain predictions for the missing data in that variable. Three regression models are considered and detailed in the appendix, and detailed in this section, the same three that are utilized in Section 3. The regression models considered are the Normal Ordinary Least Squares (OLS) regression model for a variable that is Normally distributed, the logistic Generalised Linear Model (GLM) for a variable that is dichotomous or binary in nature, and the Poisson GLM for a variable that displays count data. For more information on the ordinal Probit GLM used for non-dichotomous categorical data, see Raghunathan *et al*. (2001).

3. APPLICATION

The main focus of the application is to determine whether or not social capital, which is generally regarded as being the networks, norms and trust present in social organizations (Putnam, 1993), has influenced welfare in households after the democratization of South Africa in 1994. In this paper social capital is measured simply by whether or not a member of the household in question has a member that is part of a social network. The social networks considered include financial (burial societies and savings societies), production (farmers' associations, informal traders' associations, sewing groups, and community garden groups), cultural (sports, dance/music/singing and study groups), service (development committees, water committees and school committees), political (tribal authorities, men's groups, women's groups and youth groups), and a catch-all category for other groups.

The variables that are used in the longitudinal analysis are total years of education of the household members, household size, years of education of the household head, gender of the household head, age of the household head, location of the household (urban/rural status), social capital (networking potential), and welfare (log of real monthly adult equivalent expenditure per month without remittances). The former variables are typically significant determinants of welfare (see for example Narayan and Pritchett, 1997; Yúnez-Naude and Taylor, 2001; Grootaert *et al*., 2002; and Grootaert and Narayan, 2004), and, thus, they are a logical choice for this research.

Several regression analyses are performed using different determinants. However, the differences in determinants are only concerned with the social capital measures – the remainder of the determinants are unchanged throughout this section's analyses. Social capital is *firstly* represented separately as a single dichotomous 'access to networks' variable, then *secondly* as access to financial social capital, and *finally* as a set of dichotomous variables, each indicating 'access to a particular type of network'. The changes are made in order to gain insight into the dynamics of different types of networking social capital, although not all the results are reported.

For each set of determinants three panel regressions are performed. In the *first*, the regressions are performed on the original data, essentially dropping observations with missing values as the procedures are complete-case analysis methods. *Secondly*, the entire SRMI 'filled-in' dataset is used in the regression analysis. *Finally*, the observations with the dependent variable originally missing are dropped from the analysis and the remainder of the data is used in the regression analysis. The argument for carrying out this last procedure is that the SRMI process forces a relationship between the missing dependent data and the rest of the dataset – in other words adding these cases to the analysis may force the analyses into showing a stronger a relationship between the dependent variable and the explanatory covariates than is actually present in the population. For brevity's sake, only the first SRMI results are reported, and then only for the case when social capital is represented by the set of dummy variables for its component networks. The analysis of the dataset with SRMI filled-in values and the original incomplete dependent variable yields results similar to the case without SRMI, but these will be discussed in more detail below.

One must note that in the previous paragraphs the results from a single regression are actually created by separately regressing the dependent variable on the independent variables in each of the imputed datasets arising from the multiple imputation process. Once all the regression results are combined using Rubin's Rules, a single estimation procedure is said to have been performed. It is these combined results that are reported in the tables in this section.

In addition to the SRMI procedure detailed in the previous section, the application was extended to include an analysis of the datasets that were filled with imputations from a panel-adjusted SRMI procedure. In essence, independent variables in each sequential regression in the SRMI procedure included the filled-in variables from the previous waves as additional covariates. The results from this additional procedure and analysis were, however, so close to the original results that, for brevity's sake, they are not discussed here.

The amount of missing data that we are dealing with in this section across the three waves of the survey increases drastically across the years. Although the missing data seems to be within reasonable bounds in the first wave of the survey, more and more data is missing in the subsequent waves i.e. it changes in most cases from around 1% to 16 % to 40% over the waves. In all three waves of the survey, the variables with the most missing entries are those measured for the head of the household, namely years of education of the head of household, gender of the head of household, and age of the head of household, a missing proportion that moves from 15% to 26% to 55% over the three waves.

Another aspect that can be noted is that in 1998 the social capital data was either all missing or all present. This leads to the same amount of missing data in each social capital variable. One might argue that, if this is the case, the sequential regression imputation procedure might produce different results depending on which social capital variable was regressed first in the SRMI procedure. In order to test this argument, another multiple imputation run and panel regression was performed when all the social capital dichotomous variables were to be included as determinants (and of course only when SRMI is applied). In this additional procedure, service social capital and financial social capital were swapped in the order in which they were regressed in the SRMI process, so that service social capital would be imputed before production, cultural and financial social capital. The results appear to be similar to those where the two social capital variables are not swapped, indicating that at least using this particular set of real data, the order of the sequential regressions in the SRMI process does not influence the results in any significant way.

Complete-Case Analysis after SRMI

In each SRMI application ten iterations were performed, meaning that ten separate datasets were created for each ECM regression. After SRMI and the ECM regressions on the ten datasets were completed, the regression estimates on each of the ten imputed datasets were then combined using Rubin's Rules to yield the estimates tabulated below.

Table 1. Complete-case error component model estimation after SRMI

After SRMI	Coefficient estimates	Standard errors	Lower CI bound	Upper CI bound	Efficiency
Household education	0.0034	0.0041	-0.0058	0.0126	0.9540
Household size	-0.0278	0.0136	-0.0584	0.0029	0.9540
African household	**-0.8677**	**0.0535**	**-0.9744**	**-0.7610**	**0.9818**
Head's education	**0.0303**	**0.0109**	**0.0057**	**0.0549**	**0.9537**
Female head	**-0.1603**	**0.0420**	**-0.2491**	**-0.0716**	**0.9640**
Age of head	**0.0075**	**0.0016**	**0.0040**	**0.0110**	**0.9623**
Rural cluster	**-0.3232**	**0.0587**	**-0.4472**	**-0.1992**	**0.9639**
Financial social capital	**0.1831**	**0.0335**	**0.1148**	**0.2515**	**0.9726**
Production social capital	0.0174	0.1191	-0.2418	0.2767	0.9587
Cultural social capital	**0.2213**	**0.0568**	**0.1037**	**0.3389**	**0.9685**
Service social capital	0.1320	0.0766	-0.0276	0.2916	0.9670
Political social capital	**0.2081**	**0.0743**	**0.0537**	**0.3624**	**0.9680**
Other social capital	**0.3061**	**0.1170**	**0.0618**	**0.5505**	**0.9665**
Constant	**7.1226**	**0.1004**	**6.9083**	**7.3370**	**0.9620**

Table 2. Additional statistics for the model estimation, after SRMI

R-sq: within	0.6122
R-sq: between	0..6642
R-sq: overall	0..6700
Observations	3928
Groups	1906

In all the estimations performed, the coefficients of the final results are as we would expect them. Across the analyses the coefficient signs are identical, while the coefficient magnitudes are also similar. Household education, education of the household head, social capital, and age of the household head are all positively associated with welfare across the analyses. Additionally, larger households, African households (as opposed to Indian households), households headed by females, and rural households are associated with lower welfare levels. Note that larger households spend less per adult equivalent than smaller households do, which explains the lower welfare measure in this study.

Without SRMI, and after SRMI with the original missing values in the dependent variable still missing, almost all coefficients from the ECM regression are always significant, the exception being production social capital when social capital is represented by its individual components. When SRMI is performed and all the cases are kept in the ECM regressions, household education and size lose significance when social capital is represented by every one of its three variants: when it is a single dichotomous variable, separated into distinct dichotomous variables and when only financial social capital is entered into the analysis. This is the most important result of this study. Without SRMI, or without imputation-filled dependent variables, complete case analysis fails to recognize the relative insignificance of household education and household size. It seems that the most important seed for an increase in welfare might be an educated and senior household head, and social capital, if one ignores factors such as household head gender, race and location, which the Government will not be able to address.

The R-square statistics for the published results are quite high, especially the 'within' statistic. This can be due to the fact that the final ECM regression covariates are used as predictors of the ECM regression dependent variable in the actual SRMI process. This may not be a problem, though, since if the regression relationship uncovered in the SRMI process is assumed to be correct, the final relationship between the ECM regression dependent variable and the ECM regression independent variables is also correct.

The regression R-square statistics are around 6%, 58%, and 53% for within, between, and overall R-square, respectively, for each of the three estimations run (one for each different representation of social capital) in the case of the ECM regression analysis made before SRMI and after SRMI when the dependent variable retains its original missing data. Of course, over 850 cases are dropped without SRMI, and close to 620 are dropped if the original incomplete dependent variable is used in the ECM regression analysis after SRMI.

In all the estimations the efficiencies are always high, as in the reported estimation. This indicates robustness in the estimates, and gives strength to a pattern that will be detailed in the concluding section. Non-robust estimates lose their significance in a few of the regressions, these variables being household size and the education of the head of household.

The SRMI process itself has provided adequate results. The ten iterations that were applied in the SRMI process seem to have been enough to increase uncertainty in our estimates (to account for the uncertainty in the imputation process), and also not too few that the estimates' standard errors are too large. In fact, every

analysis reveals efficiencies of over 95%, meaning that the standard deviations of the parameter estimates are less than 1.0526 times the size of the standard errors after an infinite number of imputations.

4. CONCLUSION

It is clear from the application section that SRMI on the PSLSD/KIDS incomplete dataset has influenced the error component model parameter estimates. Before SRMI is applied to the dataset all of the covariates are significant in the model. However, once SRMI has been applied to the data several covariates lose their significance, showing less of an impact than before. It is clear that the analysis of the data is confounded when missing values are not accounted for, meaning that blocks of data that were dropped by the complete case analysis before SRMI are important to the analysis. This boils down to the missing values not being MCAR.

It must be noted again that further research into the SRMI process is warranted. Using mixed models inside SRMI process could provide more accurate predictions for missing values. Also, finding measures to correctly determine the number of rounds and iterations needed in the imputation process is of interest.

In this paper, the error component model estimates arising from the real data analysis using all of the observations are similar to those from the analysis using only the observations for which the dependent variable was originally missing. These estimates consistently show that the location of the household, the race of the household, the gender of the head of household, and social capital access are significant determinants of household welfare. Moreover, financial social capital counts for much of the influence of social capital in general, while cultural social capital also has a role to play. The most important determinant is consistently race, with location always second to this. African households suffer lower welfare than Indian households do. The same goes for rural households, who suffer poorer welfare than urban households. The results can be considered eye-opening for several more reasons, however. *Firstly*, social capital access is shown to be a more important determinant of welfare than education is, when welfare is measured by household expenditures. This raises the question of whether or not Government is promoting the formation of social capital enough. *Secondly*, the gender of the household head is consistently significant, showing that female-headed household *still* suffer poorer welfare than male-headed households do, even ten years after the start of the new South Africa. *Thirdly*, the effects of race are still very strong. While the world is striving for equality, it seems South Africa needs to be careful not to fall behind. The divides between Indian and African and between urban and rural households are extensive, even after ten years of development plans.

It must be conceded that the model for welfare is by no means complete in this paper. This fact is echoed by the poor coefficients of variation that are uncovered in the error component model estimation after SRMI (including only observations with non-missing dependent variables), coefficients of variation that are poorer even than in the original complete-case analysis. It is clear that further research is needed into finding the covariates that truly do determine household welfare. So, while this research has rejected it's hypothesis of interest – that coefficient estimates are unchanged after SRMI – it is clear that further research is required in many avenues, namely GLMMs In the SRMI process, an appraisal of the number of rounds and imputations used in the SRMI process, the (unbiased) determinants of household welfare in KwaZulu-Natal, and in social capital – is there more that South Africa could be doing to promote this useful commodity?

APPENDIX: IMPUTING FROM GENERALISED LINEAR MODELS (GLMS)

Normal Data

When the variable in question is distributed Normally, i.e. $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$, then the OLS regression model is applicable, where $\mathrm{E}[\boldsymbol{y}] = \mathrm{X}\boldsymbol{\beta}$. As noted in Section 2.3.1 a random draw from the posterior of the parameters and σ^2 is needed, and from there a random draw can be made from the predictive posterior of the variable.

The parameter estimates for OLS are easily shown to be $\widehat{\boldsymbol{\beta}} = \left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\boldsymbol{y}$. In order to generate a random draw from the posterior of σ^2 we note that:

$$U = \frac{\left(\boldsymbol{y} - \mathrm{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \mathrm{X}\widehat{\boldsymbol{\beta}}\right)}{\sigma^2} \sim \chi^2_{n-k}$$

where n is the number of observations in the regression and k is the number of parameters. The joint prior used is the Jeffrey's prior, $\frac{1}{\sigma^2}$.

Generating a random draw, u, from the χ^2_{n-k} distributions, and using the parameter estimates, $\widehat{\boldsymbol{\beta}}$, one can generate an estimate for σ^2, namely, σ^2_*, using the following equation:

$$\sigma^2_* = \frac{\left(\boldsymbol{y} - \mathrm{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \mathrm{X}\widehat{\boldsymbol{\beta}}\right)}{u}$$

Using this estimate one can draw a set of parameters, $\boldsymbol{\beta}^*$, from the posterior distribution of the parameters, using:

$$\boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}} + \mathrm{T}z_1 \tag{1.8}$$

where T is the symmetric square root of the covariance matrix and z_1 is a random draw from the Standard Normal distribution.

Using $\boldsymbol{\beta}^*$and σ^2_*, one can impute missing values using the following equation:

$$y^*_{mis} = \mathrm{X}_{mis}\boldsymbol{\beta}^* + \sigma_* z_2 \tag{1.9}$$

where z_2 is another random draw from the Standard Normal distribution.

Binary Data

When the variable in question is binary, one should implement a special case of the Binomial model, in which $\boldsymbol{y} \sim Bin(\boldsymbol{n}, \boldsymbol{\pi})$. With dichotomous data the elements of $\boldsymbol{n}$ are ones. The GLM that is used to estimate parameters for this model is the general logistic regression model, with the following link function:

$$logit(\boldsymbol{\pi}) = \ln\left(\frac{\boldsymbol{\pi}}{1 - \boldsymbol{\pi}}\right) = \mathrm{X}\boldsymbol{\beta} \tag{1.10}$$

Maximum likelihood estimates of the parameters $\boldsymbol{\beta}$, and therefore also of the vector of probabilities $\boldsymbol{\pi} = \frac{\exp(\mathrm{X}_{mis}\boldsymbol{\beta})}{1+\exp(\mathrm{X}_{mis}\boldsymbol{\beta})}$, are obtained by maximizing the following log-likelihood function:

$$l(\boldsymbol{\pi}; \boldsymbol{y}) = \sum_{i=1}^{n} [y_i \ln \pi_i + (1 - y_i)\ln(1 - \pi_i)] \tag{1.11}$$

From Equation (2.10) we have that

$$\pi_i = \frac{\exp(x_i\boldsymbol{\beta})}{1+\exp(x_i\boldsymbol{\beta})}$$

and therefore

$$\ln(\pi_i) = x_i\boldsymbol{\beta} - \ln[1+\exp(x_i\boldsymbol{\beta})]$$

and

$$\ln(1-\pi_i) = -\ln[1+\exp(x_i\boldsymbol{\beta})]$$

Using these results in Equation (2.11) yields the following log-likelihood function to be maximized:

$$l(\boldsymbol{\pi};\boldsymbol{y}) = \sum_{i=1}^{n}\{y_i[x_i\boldsymbol{\beta} - \ln[1+\exp(x_i\boldsymbol{\beta})]] - (1-y_i)\ln[1+\exp(x_i\boldsymbol{\beta})]\}$$

$$= \sum_{i=1}^{n}\{y_i x_i\boldsymbol{\beta} - \ln[1+\exp(x_i\boldsymbol{\beta})]\} \qquad (1.12)$$

For maximum likelihood estimation, the scores with respect to the $(p+1)$ elements of $\boldsymbol{\beta}$ are required, $U_0, U_1, \ldots, U_p,$, or in other words the derivatives of the log-likelihood function with respect to the elements of $\boldsymbol{\beta}$, as well as the information matrix, F. The estimates are then obtained by solving the iterative equation $\mathrm{F}^{(m-1)}\widehat{\boldsymbol{\beta}}^{(m)} = \mathrm{F}^{(m-1)}\widehat{\boldsymbol{\beta}}^{(m-1)} + \boldsymbol{U}^{(m-1)}$, where the superscripts denote the number of the iteration. The initial settings for the elements of $\widehat{\boldsymbol{\beta}}$ are zeros. Estimates are taken once convergence has been achieved, and at that stage the covariance matrix is taken as the inverse of the information matrix. For more details on the process, see Dobson (2002).

To impute missing values from this distribution, a random draw, $\boldsymbol{\beta}^*$, is drawn from the posterior of the parameters as before in Equation (2.8). Then a vector of probabilities is generated:

$$\boldsymbol{\pi}_* = \frac{\exp(\mathrm{X}_{mis}\boldsymbol{\beta}^*)}{1+\exp(\mathrm{X}_{mis}\boldsymbol{\beta}^*)}$$

Finally, a vector of Uniform random variables is generated that has the same length as $\boldsymbol{\pi}_*$, and this vector is compared with $\boldsymbol{\pi}_*$. If an element of the vector of Uniforms is less than or equal to the corresponding element of $\boldsymbol{\pi}_*$ then a '1' is imputed for the missing value associated with that element of $\boldsymbol{\pi}_*$. Alternatively, if an element of the vector of Uniforms is greater than the corresponding element of $\boldsymbol{\pi}_*$ then a '0' is imputed for the missing value associated with that element of $\boldsymbol{\pi}_*$. This process details approximate draws from the predictive posterior of the missing values (Raghunathan *et al*., 2001).

Count Data

For count data, where $\boldsymbol{y} \sim Pois(\boldsymbol{\lambda})$, the GLM to be used is the Poisson regression model, represented by the following equation:

$$\boldsymbol{\lambda} = \exp(\mathrm{X}\boldsymbol{\beta})$$

The linear predictor $g(\boldsymbol{\lambda}) = \mathrm{X}\boldsymbol{\beta}$ is used, with $g(.)$ being the log link function.

Estimation of the parameters occurs in the same way as with regular OLS estimation once the dependent variable has been transformed using the log link function.

Once more a random draw, $\boldsymbol{\beta}^*$, is taken from the posterior of the parameters of the regression model, as before in Equation (2.8). A parameter set, $\boldsymbol{\lambda}^*_{mis}$, is then generated as follows:

$$\boldsymbol{\lambda}^*_{mis} = \exp(\mathrm{X}_{mis}\boldsymbol{\beta}^*)$$

A missing datum is then imputed by drawing a random number from a Poisson distribution with the element of $\boldsymbol{\lambda}^*_{mis}$ corresponding to that missing datum as the distribution's parameter.

Imputing from Generalised Linear Mixed Models (GLMMs)

It should be mentioned that, just as the datasets created in Section 3 are analysed using linear mixed models incorporating both fixed and random effects, the creation of the datasets themselves can incorporate the same methods. In other words, the sequential regressions used in the multiple imputation process can incorporate both fixed and random effects, to be more accurate and true to the data that will be used in this research paper. However, this process of including random effects into generalised linear models is not simple when prediction is concerned. In order to predict missing values, random effects have to be estimated in each generalised linear mixed model for each group in the data (namely households in Section 3's analysis). This is not explicitly possible from model estimation, but rather requires a form of a Gibbs sampling to determine the separate random effects. It is therefore not entirely a drawback to exclude this additional complexity in the SRMI processes that follow in this research paper. Indeed, it is still worthwhile to investigate the applicability of a simpler SRMI model, against which further research into GLMM SRMI can be compared.

REFERENCES

Dobson, A. J. 2002. *An Introduction to Generalised Linear Models* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Grootaert, C., and Narayan, D. 2004. Local Institutions, Poverty and Household Welfare in Bolivia. *World Development* 32(7): 1179-1198.

Grootaert, C., Oh, G.-T., and Swamy, A. 2002. Social Capital, Household Welfare and Poverty in Burkina Faso. *Journal of African Economics* 11(1): 4-38.

KwaZulu-Natal Income Dynamics Study. 2003. *KwaZulu-Natal Income Dynamics Study (KIDS) 1993-1998. Overview of Data Files*. Release version No. 3.

Little, R. J. A., and Rubin, D. B. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Narayan, D., and Pritchett, L. 1997. *Cents and Sensibility: Household Income and Social Capital in Rural Tanzania*. Policy Research Working Paper No. 1796. Washington D.C: World Bank.

Putnam, R. D. 1993. The Prosperous Community. *American Prospect* 4(13): 4-10.

Raghunathan, T. E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. 2001. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology* 27(1):85-95.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63(3): 581-592.

Schafer, J. L. 2003. Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. *Statistica Nederlandica* 57(1): 19-35.

Schafer, J. L., and Graham, J. W. 2002. Missing Data: Our View of the State of the Art. *Psychological Methods* 7(2):147-177.

Yúnez-Naude, A., and Taylor, J.E. 2001. The Determinants of Nonfarm Activities and Incomes of Rural Households in Mexico, with Emphasis on Education. *World Development* 29(3): 561-572.

Time Series Analysis of the Southern Oscillation Index using Bayesian Additive Regression Trees

S. van der Merwe, Department of Mathematical Statistics and Actuarial Science, IB75, University of the Free State, Box 339, Bloemfontein, 9300, South Africa
October 2009

Abstract

Bayesian additive regression trees (BART) is a new regression technique developed by Chipman *et al*. (2008). Its usefulness in standard regression settings has been clearly demonstrated, but it has not been applied to time series analysis as yet. We discuss the difficulties in applying this technique to time series analysis and demonstrate its superior predictive capabilities in the case of a well know time series: the Southern Oscillation Index.

Introduction

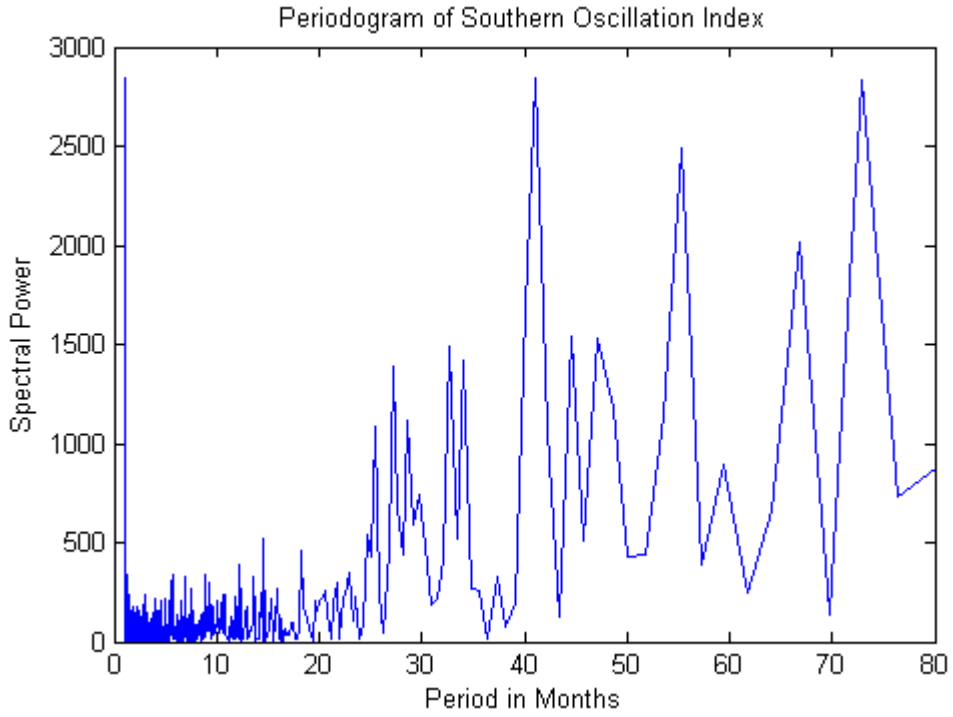
The Southern Oscillation Index (SOI) measures the difference in air pressure between Tahiti and Darwin. This value has a strong influence on weather patterns. In De Waal (2009) we see how the October SOI value in particular affects the rainfall and river flows in South Africa. Being able to accurately predict this value into the future is thus very important.

In the past the value of the SOI has been predicted using a classical ARMA approach (Chu & Katz, 1985). This approach is limited by the fact that is a linear model of the past values and errors. Non-linear models, on the other hand, present unique challenges to the researcher wishing to apply them. A good example is the use of Neural Networks for time series analysis (Giordano, *et al*. 2007).

The biggest challenge is creating prediction intervals, as the standard method does not extend to complex non-linear models. Alternatives exist, but they are computationally expensive. BART models do naturally produce a predictive posterior distribution for predictions, but extending these predictive intervals further than one time interval into the future is not a simple matter.

According to Chipman *et al*. (2008), BART models work by summing a large number of small decision trees where each tree explains a small portion of the variation in the target variable. These trees are kept small by placing a strong prior distribution on the size of each tree. The trees themselves differ from standard decision trees in that they are built randomly using a MCMC back-fitting technique. They go on to show that these models have extraordinary predictive power when compared to established techniques.

In order to predict the SOI it may be useful to gain a deeper understanding of the properties of the series. It seems to be a naturally oscillating series with multiple long term cycles, as seen by the extreme peaks in the periodogram below. It is worth investigating the effect of including these cycles in any model one may fit to this series.



The values of the SOI that were used in this analysis were the values from January 1876 until August 2009, making this series fairly long by time series standards. However, BART is by its very nature a data mining model that works best when presented with very large samples. It is this nature that also makes these models prone to over-fitting. We thus treat the modelling of the series using BART as a data mining problem but randomly dividing the sample into a training portion and a testing portion.

Regression Approach to Predicting October's SOI Value

Calculation of Model Accuracy

Various measures of model accuracy were calculated by comparing the predictions with the observed values for all past Octobers, but distinguishing between those values that were used to fit the model (sample values) and those values that were kept aside (out-of-sample values). Three of these are reported here for every model: the correlation coefficient (CORR), the mean absolute error (MAE) and the root mean square error (RMSE).

It is important to note that BART models are random models, in that they produce different results every time they are fitted. Also, the allocation of observations to the sample versus out-of-sample is random. Thus, all results and measures reported are an average taken over ten fits, or runs, of the model.

Using all available variables

If we build a model to predict October's value using the twelve months prior as well the as 41st and 73rd months prior to each October we notice the presence of over-fitting, that is to say that the out-of-sample predictions are significantly weaker than the sample predictions.

The model mentioned above with 14 explanatory variables produces the following mean fit statistics:

Sample	CORR	MAE	RMSE
Training	0.92	3.23	4.04
Testing	0.72	4.89	6.18

Note that for all models in this section a random 80% of the observations are used to train the model and the remaining 20% are used for testing.

Variable Selection

One of the reasons BART models over-fit is because too many trees are used, but that is not the case here, as we have reduced the number of trees to just twenty. In fact, the number of trees used does not seem to have much impact on the results at all. The results are only slightly weaker when say ten or forty trees are used.

The reason for overfitting here seems to be that too many variables are included in the model. We need to perform some form of variable selection, but we need to bear in mind that we are working with a heavily non-linear model and that most standard variable selection techniques are not appropriate. Thankfully, BART has its own, unique, method of variable selection.

It works by reducing the number of trees to the point where the model has to be selective about which variables to use in each branch. The model is then forced to seek out those variables that are most useful in predicting the target variable. Counting the number of times each variable is used in the model produces a measure of the relative importance of each variable.

Applying this technique to our model produces the following relative variable importance (1 = medium relative importance):

Variable	Sep	Aug	Jul	Jun	May	Apr	Mar
Importance	3.04	1.21	1.68	1.35	1.00	0.60	0.49
Variable	Feb	Jan	Dec-1	Nov-1	Oct-1	-41	-73
Importance	0.67	0.42	0.96	0.58	0.44	0.93	0.62

Reduced Model

Based on this we thus reduce our choice of variables to only the five months prior to each October. This reduces the overfitting problem and gives results as follows:

Sample	CORR	MAE	RMSE
Training	0.90	3.45	4.33
Testing	0.81	4.89	5.90

Note that these results can be improved by fitting the model multiple times and selecting the best model.

There is, however, one glaring problem with this model: we only gain one month by using it. We must have September's value in order to predict October's value.

Excluding September

We can gain another month by excluding September for the previous model. This model is slightly weaker but is, nevertheless, the best model we can produce under these constraints for the purpose of predicting the October value. Its fit statistics are as follows:

Sample	CORR	MAE	RMSE
Training	0.82	4.49	5.60
Testing	0.66	5.65	7.16

Predictions from further back in time

Making predictions using only higher lags is very difficult. We have already seen that there is some relationship with the previous December as well as May of three years

prior but we were unable to find additional useful variables. Thus, we can, at best, achieve results as follows:

Sample	CORR	MAE	RMSE
Training	0.49	7.10	8.58
Testing	0.22	7.80	9.40

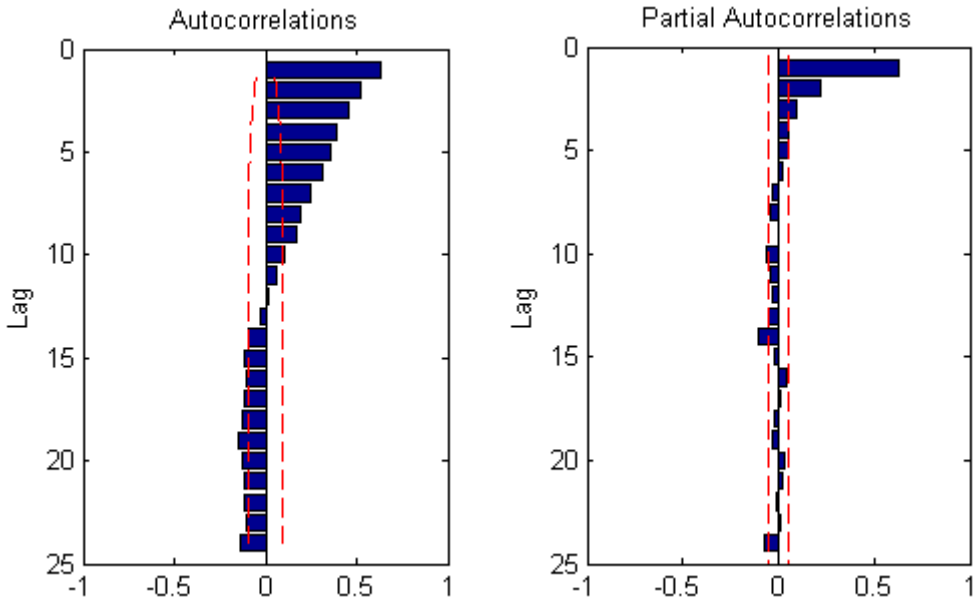
It is possible to achieve better predictions for October by not focusing on October itself. This is especially true in the first half of each year.

Autoregressive Time Series Approach

In this section we consider every month equal and try to make predictions for the months ahead, completely ignorant of the current date. This is closer to the classic univariate time series approach.

When we use this approach we have twelve times as many observations and so over-fitting is less of a concern (but still worth baring in mind). Here we increase the number of trees in each model to 40. We also keep aside a random third of the observations for testing (as opposed to a fifth).

Again, we need to determine the current variables to include in the model. In time series analysis, one method that is often used is the inspection of the correlograms. If we look at the correlograms below and we restrict ourselves to autoregressive models then it is clear that an AR(5) model is appropriate.



However, autocorrelations are a linear measure of dependence, and so we once again apply BART variable selection:

Lag	1	2	3	4	5	6
Importance	4.71	2.11	1.22	0.47	0.83	0.39
Lag	7	8	9	41	73	
Importance	0.19	0.29	0.21	0.42	0.17	

From the table it is clear that using the first 5 lags is, in fact, appropriate. Making predictions into the future using BART models of this form gives the following out-of-sample fit statistics:

Out of Sample	**Correlation**	**MAE**	**RMSE**
1 Month	0.647	6.158	7.886
2 Months	0.552	6.763	8.624
3 Months	0.483	7.083	9.069
4 Months	0.425	7.365	9.381
5 Months	0.385	7.554	9.555
6 Months	0.346	7.697	9.710
7 Months	0.304	7.852	9.880
8 Months	0.253	7.980	10.041
9 Months	0.239	8.001	10.076
10 Months	0.220	8.052	10.122
11 Months	0.233	7.967	10.069
12 Months	0.218	8.014	10.121

In the table above the predictions are made using the mean of the posterior distribution.

If we use the median of the predictive posterior instead then we get worse results:

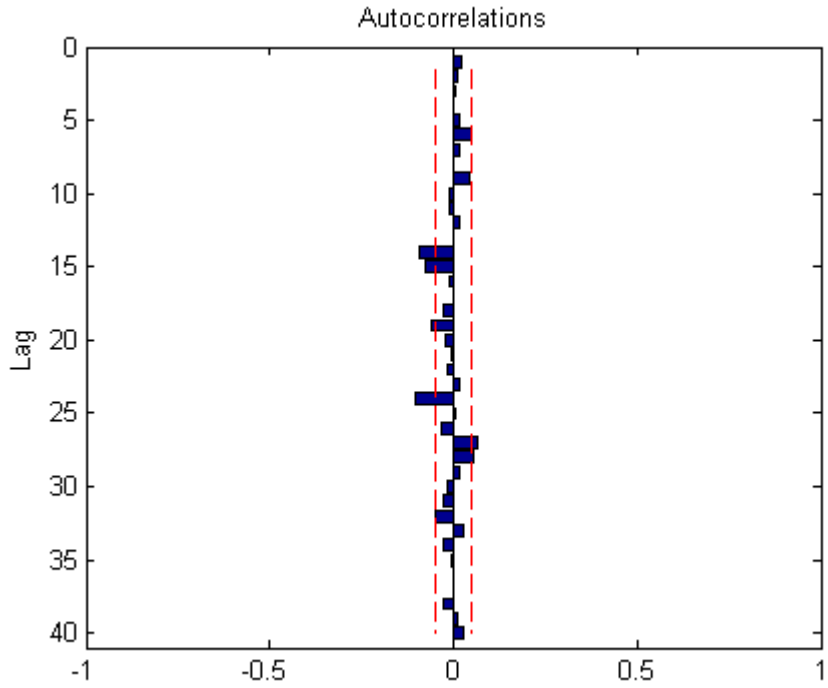
Out of Sample	Correlation	MAE	RMSE
1 Month	0.638	6.238	7.978
2 Months	0.542	6.849	8.703
3 Months	0.477	7.130	9.114
4 Months	0.420	7.420	9.417
5 Months	0.374	7.651	9.633
6 Months	0.325	7.779	9.825
7 Months	0.281	7.918	9.975
8 Months	0.243	8.024	10.094
9 Months	0.207	8.110	10.193
10 Months	0.203	8.084	10.171
11 Months	0.198	8.093	10.197
12 Months	0.193	8.093	10.202

Making predictions into the future with BART models is a great deal more difficult than it is with standard time series models as BART models do not currently allow for situations where the output is to be fed back through the model as input.

If we are merely interested in a point estimate (as above) then we can take the mean prediction as fact and re-fit the model with this value added onto the end of the series. This produces a prediction for two months ahead of the current month, which can, in turn, be added onto the series, *etc*.

If we require predictive intervals then the above process must be repeated many times using a random value from the predictive posterior distribution instead of the mean or median prediction. This process requires a distributed computing environment to avoid excessive run times on current generation computing hardware.

It is worth noting that this model produces white noise residuals when making one-month-ahead predictions. This can be seen quite clearly in the correlogram below:



Comparison and Conclusion

If we fit a linear AR5 model in the same way we obtain the following out-of-sample fit statistics:

Out of Sample	Correlation	MAE	RMSE
1 Month	0.620	6.154	7.937
2 Months	0.532	6.743	8.564
3 Months	0.422	7.207	9.197
4 Months	0.374	7.383	9.402
5 Months	0.338	7.466	9.524
6 Months	0.261	7.676	9.815
7 Months	0.171	7.879	10.090
8 Months	0.156	7.807	10.059
9 Months	0.163	7.705	9.991
10 Months	0.164	7.745	9.980
11 Months	0.148	7.761	10.016
12 Months	0.112	7.784	10.080

This is worse than the BART model by some way. However, if we look at the BART predictions for the coming months we encounter a problem:

Sep 09	Oct 09	Nov 09	Dec 09	Jan 10	Feb 10
-1.97	0.40	0.77	0.09	1.15	0.71
Mar 10	Apr 10	May 10	Jun 10	Jul 10	Aug 10
0.76	0.66	0.61	0.75	0.73	0.67

It appears as though the model gradually tends towards a flat model with the variance fading after 5 months. This may produce a relatively good fit in statistical terms but it is of little value to the researcher wishing to know what future values of the series are likely to be.

We obtain similar (but worse) results from the AR(5) model:

Sep 09	Oct 09	Nov 09	Dec 09	Jan 10	Feb 10
0.27	0.55	0.06	0.04	0.15	0.10
Mar 10	Apr 10	May 10	Jun 10	Jul 10	Aug 10
0.09	0.07	0.05	0.04	0.03	0.02

In conclusion, the SOI remains a very difficult time series to predict, in spite of the power of BART models.

References

Chipman, H.A., George, E.I. and McCulloch, R.E.: 2008, Bart: Bayesian additive regression trees. *http://arxiv.org/abs/0806.3286*

Chu, P.S., Katz, R.W.: 1985, Modelling and forecasting the southern oscillation: A time-domain approach, *Monthly Weather Review* **113**(11),pp.1876-1888.

De Waal, D.J.: 2009, Predicting losses due to spillage at the Gariep hydro-power plant. Report to ESKOM, Feb 2009.

Giordano, F., Rocca, M.L., Perna, C.: 2007, Forecasting nonlinear time series with neural network sieve bootstrap, *Computational Statistics & Data Analysis* **51**,pp.3871-3884.

Southern Oscillation Index: 2009, National Climate Centre, Australian Bureau of Meteorology. *ftp://ftp.bom.gov.au/anon/home/ncc/www/sco/soi/soiplaintext.html*

Biased reduced estimators in joint tail modelling

Beirlant J.[a], Dierckx G.[a,b], Guillou A.[c]

[a] Department of Mathematics and Leuven Statistics Research Centre, Katholieke Universiteit Leuven

[b] Department of Mathematics and statistics, Hogeschool-Universiteit Brussel

[c] IRMA, Département de Mathématique, Université de Strasbourg, France

October 1, 2009

Abstract

In this article we study the bias properties of the estimators based on the first order model presented by Ledford and Tawn (1997). These authors introduce the coefficient of tail dependence η, giving information about the dependence of the extreme values of two variables. We propose a bias reduced estimator for this coefficient and show its properties via simulations and real life examples. Further on, a bias reduced estimator for small tail probabilities follows immediately.

Keywords: Extreme value theory, bivariate slowly varying function, bias reduction

1 Joint tail modelling

Let (Z_1, Z_2) be a bivariate stochastic random variable with marginal distributions which have already been studied. Then, without loss of generality, we can assume that Z_1 and Z_2 have Fréchet margins. Indeed, for bivariate (X_1, X_2) with general margins, the data can be transformed to Fréchet marginals using the empirical distribution functions $\widehat{F}_{X_1}(x)$ and $\widehat{F}_{X_2}(x)$.

$$\begin{aligned} Z_1 &= -1/\log \widehat{F}_{X_1}(X_1) \\ Z_2 &= -1/\log \widehat{F}_{X_2}(X_2) \end{aligned}$$

Then the dependence structure between the Fréchet margins (Z_1, Z_2) can be studied using the **model of Ledford and Tawn** (1997) as introduced in Section 9.5.3 in Beirlant et al. (2005).

$$\mathbb{P}\left(Z_1 > z_1, Z_2 > z_2\right) = z_1^{-c_1} z_2^{-c_2} \mathcal{L}(z_1, z_2), \text{ with } c_1 + c_2 = \frac{1}{\eta}. \tag{1}$$

The parameter η is called the **coefficient of tail dependence** whereas the function $\mathcal{L}$ is a **bivariate slowly varying function**. This means that there exists a function g such that

$$\lim_{t\to\infty} \frac{\mathcal{L}(tz_1, tz_2)}{\mathcal{L}(t,t)} = g(z_1, z_2). \tag{2}$$

We will assume that the function g is **homogenous of order 0**. This means by definition that $g(tz_1, tz_2) = g(z_1, z_2)$. It implies that there exists a function g^*, such that

$$g(z_1, z_2) = g^*(z_1/(z_1 + z_2)), \tag{3}$$

In other words the function g does not change on a radius.

In Section 2 the model (1) is extended introducing a second order term. Some examples are given to shown that this extended model is quite natural. This second order model will be used in order to find bias reduced estimators of the parameter models and of small probabilities in Section 3. Then in Section 4, the newly introduced estimators are studied through a simulation study. Finally some real life examples are given in Section 5.

2 Second order model

2.1 Second order condition on $\mathcal{L}$

To study how fast the limit in (2) is attained, we need a proper second order condition in the bivariate case. Therefore we try to extend well-known conditions from the univariate case. Remark indeed that the bivariate slowly varying functions are an extension of univariate slowly varying functions ℓ. Indeed the univariate homogenous functions of order 0, are the functions g such that $g(tz) = g(z)$, being the constant functions. And from the definition $\lim_{t\to\infty} \frac{\ell(tz)}{\ell(t)} = g(x) = a$, it follows that this constant must be $g(1)$=1, leading

indeed to the definition of univariate slowly varying functions.

In order to study bias properties in the univariate case, second order conditions are imposed on the univariate slowly varying function. One of these conditions is the assumption $\mathcal{R}_\ell(b,\rho)$, assuming that there exists $\rho<0$ and $b(t)\to 0$ as $t\to\infty$, such that for all $z>1$

$$\frac{\ell(tz)}{\ell(t)}\sim 1+b(t)\frac{z^\rho-1}{\rho} \text{ as } t\to\infty. \tag{4}$$

A more specific condition, is the Hall condition, where

$$\ell(t)\sim C\left(1+Dt^\rho\right) \text{ as } t\to\infty. \tag{5}$$

An extension of this Hall condition to the bivariate case could be made by replacing the constants in the univariate Hall condition again by homogenous functions of order 0. This leads to the following **bivariate Hall condition**

$$\mathcal{L}(z_1,z_2)\sim g_1(z_1,z_2)\left(1+g_2(z_1,z_2)z_1^{\rho_1}z_2^{\rho_2}\right) \text{ as } z_1,z_2\to\infty. \tag{6}$$

We will denote $\rho_1+\rho_2$ by ρ. Remark the analogy with model (2.2) in Ledford and Tawn (1997).

2.2 Examples

Remark that assumption (6) holds for a lot of well-known multivariate examples. Below we assume Fréchet margins.

2.2.1 Morgenstern

$$\mathbb{P}(Z_1\leqslant z_1,Z_2\leqslant z_2) = F(z_1)F(z_2)\left[1+\alpha\overline{F}(z_1)\overline{F}(z_2)\right]$$

$$\begin{aligned}\mathbb{P}(Z_1>z_1,Z_2>z_2) &= 1-F(z_1)-F(z_2)+P(Z_1\leqslant z_1,Z_2\leqslant z_2)\\ &= 1-e^{-1/z_1}-e^{-1/z_2}+e^{-1/z_1}e^{-1/z_2}\left[1+\alpha\left(1-e^{-1/z_1}\right)\left(1-e^{-1/z_2}\right)\right]\\ &\sim \frac{1}{z_1z_2}\left[1+\alpha-\frac{1+3\alpha}{2}\left(\sqrt{\frac{z_2}{z_1}}+\sqrt{\frac{z_1}{z_2}}\right)\frac{1}{\sqrt{z_1z_2}}\right]\end{aligned}$$

where the last steps follows from Taylor expansions to the third order of e^{-1/z_1}.

Assumption (6) holds with $c_1 = c_2 = 1$ (thus $\eta = 0.5$) and $g_1(z_1, z_2) = 1 + \alpha$. Further, $\rho_1 = \rho_2 = -0.5$ (thus $\rho = -1$) and $g_2(z_1, z_2) = -\frac{1+3\alpha}{2(1+\alpha)} \left(\sqrt{\frac{z_2}{z_1}} + \sqrt{\frac{z_1}{z_2}} \right)$.

2.2.2 Extreme Value

$$\mathbb{P}(Z_1 \leqslant z_1, Z_2 \leqslant z_2) \;=\; e^{-V(z_1,z_2)}$$

$$\begin{aligned} \mathbb{P}(Z_1 > z_1, Z_2 > z_2) &= e^{-V(z_1,z_2)} - F(z_1) - F(z_2) + 1 \\ &\sim \frac{1}{\sqrt{z_1 z_2}} \left[g_1(z_1, z_2) \left(1 + g_2(z_1, z_2) \frac{1}{\sqrt{z_1 z_2}} \right) \right], \end{aligned}$$

with

$$\begin{aligned} g_1(z_1, z_2) &= \sqrt{\frac{z_2}{z_1}} + \sqrt{\frac{z_1}{z_2}} - \sqrt{z_1 z_2} V(z_1, z_2) \\ g_2(z_1, z_2) &= -\frac{1}{2} \frac{\frac{z_2}{z_1} + \frac{z_1}{z_2} - z_1 z_2 V^2(z_1, z_2)}{\sqrt{\frac{z_2}{z_1}} + \sqrt{\frac{z_1}{z_2}} - \sqrt{z_1 z_2} V(z_1, z_2)} \end{aligned}$$

Assumption (6) holds with $c_1 = c_2 = 0.5$ (thus $\eta = 1$) and $g_1(z_1, z_2)$ as above. Further, $\rho_1 = \rho_2 = -0.5$ (thus $\rho = 1$) and $g_2(z_1, z_2)$ as above.

2.2.3 Clayton-upper tail

$$\begin{aligned} \mathbb{P}(Z_1 > z_1, Z_2 > z_2) &= \left[\left(\overline{F}(z_1) \right)^{-1/\alpha} + \left(\overline{F}(z_2) \right)^{-1/\alpha} - 1 \right]^{-\alpha} \\ \mathbb{P}(Z_1 > z_1, Z_2 > z_2) &\sim \frac{1}{\sqrt{z_1 z_2}} g_1(z_1, z_2) \left[1 + g_2^a(z_1, z_2) \frac{1}{\sqrt{z_1 z_2}} + g_2^b(z_1, z_2) \frac{1}{(z_1 z_2)^{\frac{1}{2\alpha}}} \right] \end{aligned}$$

with

$$
\begin{aligned}
g_1(z_1,z_2) &= \left[\left(\frac{z_2}{z_1}\right)^{-\frac{1}{2\alpha}} + \left(\frac{z_1}{z_2}\right)^{-\frac{1}{2\alpha}}\right]^{-\alpha} \\
g_2^a(z_1,z_2) &= -\frac{1}{2}\frac{\left(\frac{z_2}{z_1}\right)^{-\frac{1}{2\alpha}+\frac{1}{2}} + \left(\frac{z_1}{z_2}\right)^{-\frac{1}{2\alpha}+\frac{1}{2}}}{\left(\frac{z_2}{z_1}\right)^{-\frac{1}{2\alpha}} + \left(\frac{z_1}{z_2}\right)^{-\frac{1}{2\alpha}}} \\
g_2^b(z_1,z_2) &= \frac{\alpha}{\left(\frac{z_2}{z_1}\right)^{-\frac{1}{2\alpha}} + \left(\frac{z_1}{z_2}\right)^{-\frac{1}{2\alpha}}}
\end{aligned}
$$

Assumption (6) holds with $c_1 = c_2 = 0.5$ and $g_1(z_1, z_2)$ as above. Further,

$$
\begin{aligned}
&\text{if } \alpha < 1 \quad , \quad \rho_1 = \rho_2 = -0.5 \quad g_2(z_1,z_2) = g_2^a(z_1,z_2) \\
&\text{if } \alpha = 1 \quad , \quad \rho_1 = \rho_2 = -0.5 \quad g_2(z_1,z_2) = g_2^a(z_1,z_2) + g_2^b(z_1,z_2) \\
&\text{if } \alpha > 1 \quad , \quad \rho_1 = \rho_2 = \frac{1}{2\alpha} \quad g_2(z_1,z_2) = g_2^b(z_1,z_2)
\end{aligned}
$$

Note that for $\alpha > 1$, the second order term in (6) disappears slowly. This is a case where one might expect a significant improvement of the estimation when the second order term is added.

2.2.4 Bivariate normal

Due to the analogy with model (2.2) in Ledford and Tawn, we can use their results to see that $c_1 = c_2 = \frac{1}{1+\rho}$ and $\rho_1 = \rho_2 = 0$. Also here the second order term might result in improved estimators.

2.3 Model of Ledford and Tawn revisited, using second order condition

We study the following transformation of (Z_1, Z_2), based on a treshold t and a parameter w. Consider Y given by the $\min(Z_1/t, Z_2 w/[t(1-w)])$ conditioned on the fact that $\min(Z_1, Z_2 w/(1-w)) > t$. The parameter w can be chosen freely, while t can be chosen as the $k+1^{th}$ largest order statistic of $\min(Z_1, Z_2 w/(1-w))$, as is often done in extreme value statistics.

It follows from (1) combined with (6) that

$$\begin{aligned}\mathbb{P}(Y>z) &= \frac{\mathbb{P}(Z_1>zt, Z_2>zt(1-w)/w)}{\mathbb{P}(Z_1>t, Z_2>t(1-w)/w)}, \quad t>1\\ &= z^{-(c_1+c_2)}\frac{\mathcal{L}(tz, tz(1-w)/w)}{\mathcal{L}(t, t(1-w)/w)}\\ &= z^{-1/\eta}\frac{1+g_2\left(1,\frac{1-w}{w}\right)(tz)^{\rho_1+\rho_2}\left(\frac{1-w}{w}\right)^{\rho_2}}{1+g_2\left(1,\frac{1-w}{w}\right)t^{\rho_1+\rho_2}\left(\frac{1-w}{w}\right)^{\rho_2}}.\end{aligned}$$

In the last step the homogeneity of the functions g_1 and g_2 is used.
We can reparametrize this distribution as follows.

$$\mathbb{P}(Y>z) \sim z^{-1/\eta}\left(1+\delta-\delta z^{\rho}\right)^{-1/\eta}, \ t\to\infty \tag{7}$$

with $\delta := \eta g_2\left(1,\frac{1-w}{w}\right)\left(\frac{1-w}{w}\right)^{\rho_2}t^{\rho_1+\rho_2}$.

This model provides us with bias reduced estimators for η as is shown in Section 3. Moreover, the model inference can be linked to tail probabilities concerning Z_1 and Z_2. Indeed, the probability $\mathbb{P}(Z_1>z_1, Z_2>z_2)$ can be rewritten as

$$\frac{\mathbb{P}\left(Z_1>tz, Z_2>tz\frac{1-w}{w}\right)}{\mathbb{P}\left(Z_1>t, Z_2>t\frac{1-w}{w}\right)}\mathbb{P}\left(Z_1>t, Z_2>t\frac{1-w}{w}\right) \tag{8}$$

$$= \mathbb{P}(Y>z)\mathbb{P}\left(Z_1>t, Z_2>t\frac{1-w}{w}\right), \tag{9}$$

with $tz=z_1$ and $w=z_1/(z_1+z_2)$ and t chosen, as before, as the $k+1^{th}$ largest order statistic of $\min(Z_1, Z_2w/(1-w))$.

3 Biased reduced estimators

First, we recall in short some first order biased estimators in Section 3.1. Then, we turn to estimators with reduced bias in Sections 3.2 and 3.3 based on the second order model introduced in Section 2.

3.1 Biased estimation

Based on the first order terms in model (1), it immediately follows that

$$\mathbb{P}(\min(Z_1, Z_2) > r) = \mathbb{P}(Z_1 > r, Z_2 > r) = r^{-1/\eta}\ell(r),$$

with $\ell(r)$ slowly varying. In other words, the variable $\min(Z_1, Z_2)$ is Pareto-type distributed. Therefore, the Hill estimator performed on this variable provides an estimator of the parameter η. It should be remarked that the Hill estimator is based on the $k+1$ largest data of $\min(Z_1, Z_2)$. Therefor the estimator will be denoted by $\widehat{\eta}_{H,k}$.

A first order estimator for the small tail probabilities can be found from (9). Now only $\mathbb{P}(Y > z)$ is estimated by solving the Weismann estimator for large quantiles of Y for small probabilities. This estimator will be denoted by $\widehat{p}_{H,k}$. Now k refers to the treshold used for $\min(Z_1, Z_2 w/(1-w))$ in Section 2.3.

3.2 Bias reduced estimation of model parameters

Based on the tail probability in (7), maximum likelihood estimators can be determined.
Indeed, the loglikelihood of the model in (7) is given by

$$\log L = -\log\eta - (1/\eta + 1)\log z - (1/\eta + 1)\log(1 + \delta - \delta z^{\rho}) + \log(1 + \delta - \delta z^{\rho} - \delta\rho z^{\rho})$$

with partial derivatives

$$\begin{aligned}
\frac{\delta \log L}{\delta \eta} &= -\frac{1}{\eta} + \frac{1}{\eta^2}\log z + \frac{1}{\eta^2}\log(1 + \delta - \delta z^{\rho}) \\
\frac{\delta \log L}{\delta \delta} &= -(1/\eta + 1)\,\frac{1 - z^{\rho}}{1 + \delta - \delta z^{\rho}} + \frac{1 - z^{\rho} - \rho z^{\rho}}{1 + \delta - \delta z^{\rho} - \delta\rho z^{\rho}} \\
\frac{\delta \log L}{\delta \rho} &= (1/\eta + 1)\,\frac{\delta z^{\rho}\log z}{1 + \delta - \delta z^{\rho}} - \frac{\delta z^{\rho}\,(1 + (1+\rho)\log z)}{1 + \delta - \delta z^{\rho} - \delta\rho z^{\rho}}
\end{aligned}$$

In the estimation procedure however we have taken ρ equal to the canonical value -1. It is known in extreme value theory that a misspecification of the parameter ρ might not affect the estimation of the parameter of main interest that much, in this case η. We also studied the results when ρ was estimated using the method suggested by Fraga Alves et al. (2003). Indeed, the results were not much different. The resulting maximum likelihood estimator, with

ρ taken equal to -1, will be denoted by $\widehat{\eta}_{B,k}$. Again, k in the notation stresses the fact that the estimator depends on the treshold used for $\min(Z_1, Z_2 w/(1-w))$ in Section 2.3.

3.3 Bias reduced estimation of small probabilities

Bivariate tail probabilities can now be calculated using expression (9) The probability $\mathbb{P}(Z_1 > z_1, Z_2 > z_2)$ can then be estimated as follows. $\mathbb{P}(Y > z)$ can be estimated by plugging in the parameter estimates in (7) and $\mathbb{P}\left(Z_1 > t, Z_2 > t\frac{1-w}{w}\right)$ can be estimated using the empirical distribution function. The resulting estimator will be denoted by $\widehat{p}_{B,k}$.

4 Simulation results

In this section, we illustrate the finite sample properties of the previous estimators through simulation. First, we simulated from a Morgenstern with Fréchet marginals and $\alpha = 0.6$. Remark that the true values $\eta = 0.5$, $\rho = -1$. The effect of the choice of w is examined by choosing different values. The corresponding tail probabilities are estimated as well.
Next, we simulated from a normal copula with Fréchet marginals and $\rho = -0.5$. w is chosen 0.5.
Finally, we also show how the procedure works when general margins are given. We simulate from bivariate normal data (with normal marginals, instead of Fréchet marginals). Again ρ is -0.5 in the simulation.

For each setting, 10 data sets of size $n = 500$ are simulated.

4.1 Morgenstern α=0.6

For the estimation of η, the parameter w can be chosen freely. First it is taken as $w = 0.5$. In Figure 1 the estimators η_H and the biased reduced estimator η_B are compared.
Figure 1 clearly shows that η_B has better bias properties than η_H. Moreover, the choice of k seems to be less crucial for η_B.
The same conclusion can be made when estimating the probability $\mathbb{P}(Z_1 > 20, Z_2 > 20)$. The true value of the probability is 0.0036699. In this estimation in Figure 2 the value $w = 0.5$ is natural.

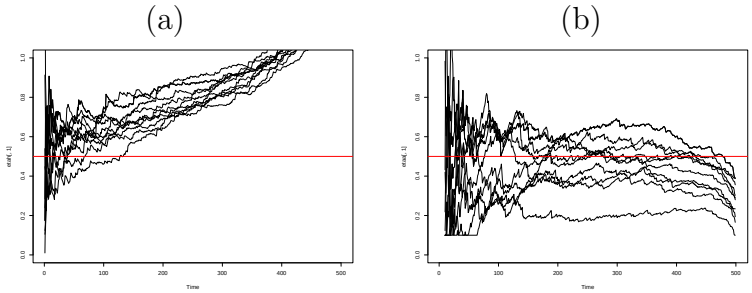


Figure 1: Estimators for η (a) $(k, \widehat{\eta}_{H,k})$; (b) $\widehat{\eta}_B$. (w=0.5)

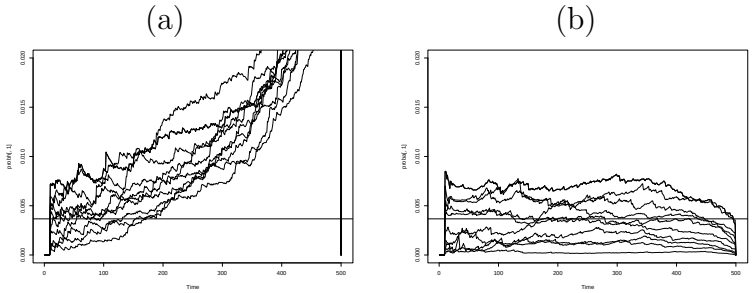


Figure 2: Estimators for $\mathbb{P}(Z_1 > 20, Z_2 > 20)$ (a) $(k, \widehat{p}_{H,k})$; (b) $\widehat{p}_B$.

The above conclusions do not heavily depend on the choice $w = 0.5$ as is illustrated in Figure 3 and Figure 4 for $w = 15/35 = 0.42857$ resp. $w = 8/28 = 0.285$

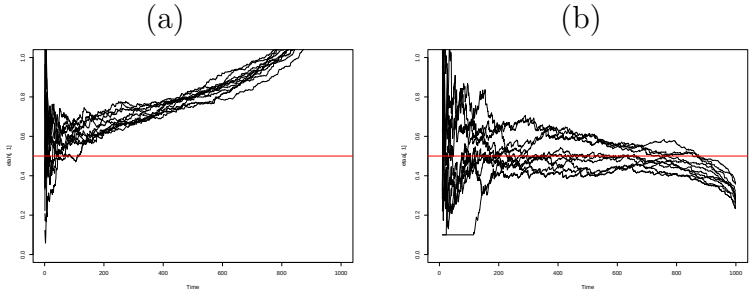


Figure 3: Estimators for η (a) $(k, \widehat{\eta}_{H,k})$; (b)$(k, \widehat{\eta}_{B,k})$. (w=15/35)

The choices of $w = 15/35$ and $w = 8/28$ are also natural choice to estimated some probabilities. The probability $\mathbb{P}(Z_1 > 15, Z_2 > 20)$ is estimated in Figure 5. The true value of this probability is 0.0048. The probability

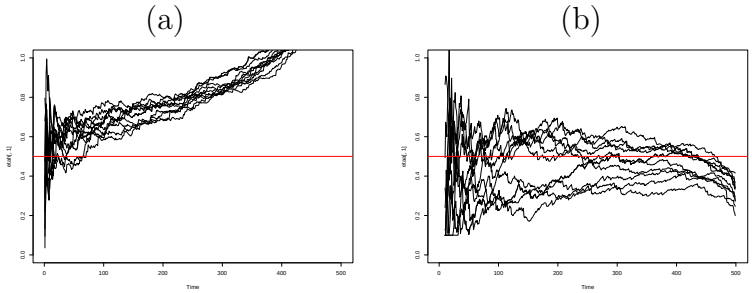


Figure 4: Estimators for η (a) $(k, \widehat{\eta}_{H,k})$; (b)$(k, \widehat{\eta}_{B,k})$. $(w = 8/28$

$\mathbb{P}(Z_1 > 8, Z_2 > 20)$ is estimated in Figure 6. Now , the true value of this probability is 0.0086.

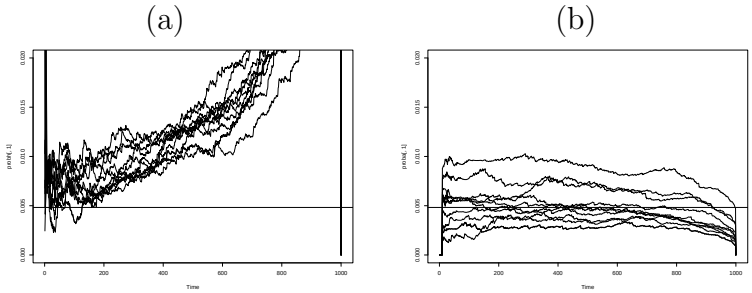


Figure 5: Estimators for $\mathbb{P}(Z_1 > 15, Z_2 > 20)$ (a) $(k, \widehat{p}_{H,k})$; (b) $(k, \widehat{p}_{B,k})$.

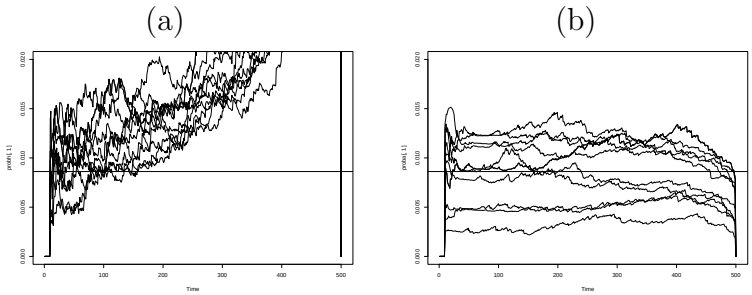


Figure 6: Estimators for $\mathbb{P}(Z_1 > 8, Z_2 > 20)$ (a) $(k, \widehat{p}_{H,k})$; (b) $(k, \widehat{p}_{B,k})$.

4.2 Bivariate normal ρ=-0.5, with Fréchet marginals

The true value of η equals 0.25. The parameter can be chosen freely and is taken $w = 0.5$ in the estimation of η in Figure 11.

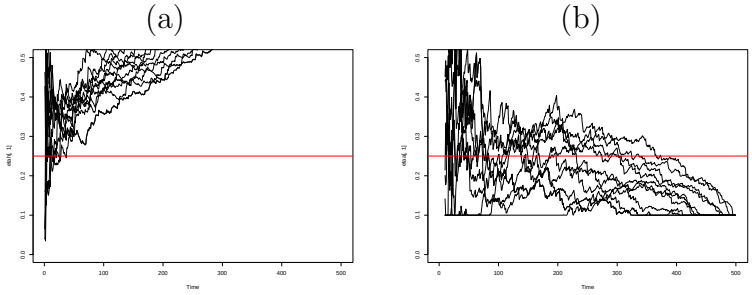


Figure 7: Estimators for η (a) $(k, \widehat{\eta}_{H,k})$; (b)$(k, \widehat{\eta}_{B,k})$.

The true value of the probability $\mathbb{P}(Z_1 > 5, Z_2 > 5)$ (natural choice of $w = 0.5$) is 0.00601.

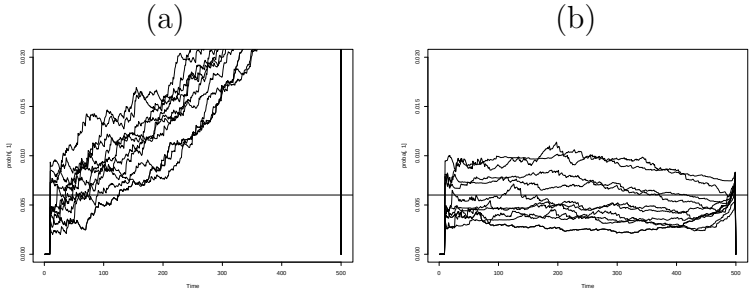


Figure 8: Estimators for $\mathbb{P}(Z_1 > 5, Z_2 > 5)$ (a) $(k, \widehat{p}_{H,k})$; (b) $(k, \widehat{p}_{B,k})$.

4.3 Bivariate normal $\rho = -0.5$, with normal marginals

First the bivariate normal data are transformed into frechet marginals :

$$\begin{aligned} Z_1 &= -1/\log \widehat{F}_{X_1}(X_1) \\ Z_2 &= -1/\log \widehat{F}_{X_2}(X_2). \end{aligned}$$

Then η can be estimated as before based on this new sample for (Z_1, Z_2). The probability $\mathbb{P}(X_1 > x_1, X_2 > x_2)$ equals

$$\mathbb{P}\left(Z_1 > -1/\log \widehat{F}_{X_1}(x_1), Z_2 > -1/\log \widehat{F}_{X_2}(x_2)\right) := \mathbb{P}(Z_1 > z_1, Z_2 > z_2)$$

which can be estimated as before.
The true value of η is 0.25. This is estimated in Figure using the choice of $w = z_1/(z_1 + z_2)$.

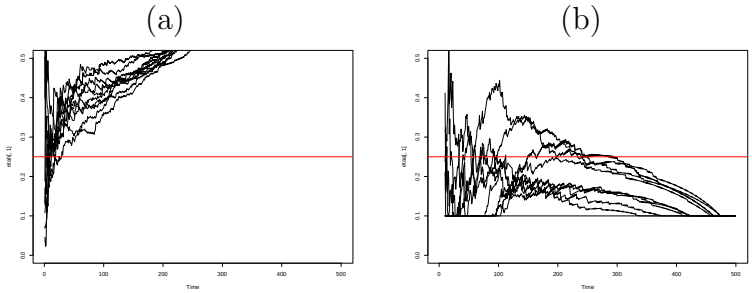


Figure 9: Estimators for η (a) $(k, \widehat{\eta}_{H,k})$; (b)$(k, \widehat{\eta}_{B,k})$.

The true value of the probability $\mathbb{P}(X_1 > 6, X_2 > 6)$ is 0.00378 with a natural choice $w = z_1/(z_1 + z_2)$. Figure shows the estimators $\widehat{p}_H$ and $\widehat{p}_B$.

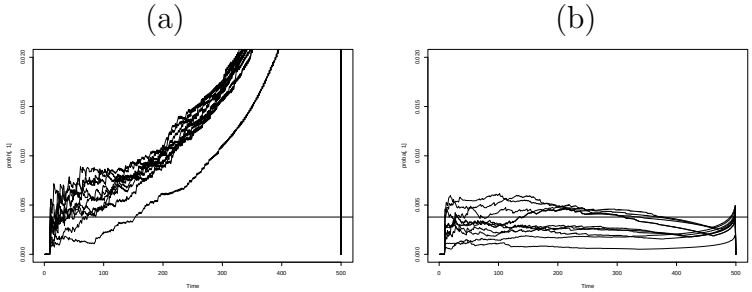


Figure 10: Estimators for $\mathbb{P}(Z_1 > 6, Z_2 > 6)$ (a) $(k, \widehat{p}_{H,k})$; (b) $(k, \widehat{p}_{B,k})$.

5 Real life examples

5.1 Data set Loss and ALAE

Consider the example of loss-ALAE data examined by Frees and Valdez (1998) and Klugman and Parsa (1999). This database records medical claim amounts exceeding 25,000 dollar. In insurance one is interested in the individual losses and the expenses that are specifically attributable to the settlement of such claims such as lawyers' fees and claims investigation expenses, abbreviated by ALAE's.

Again first the bivariate data are transformed into Fréchet marginals (Z_1, Z_2),using the empirical distribution function. To estimate $\mathbb{P}(X_1 > 200000, X_2 > 100000)$, it is transformed into Fréchet marginals to become $\mathbb{P}(Z_1 > z_1, Z_2 > z_2)$ which can be estimated as before. We choose $w = z_1/(z_1 + z_2)$.

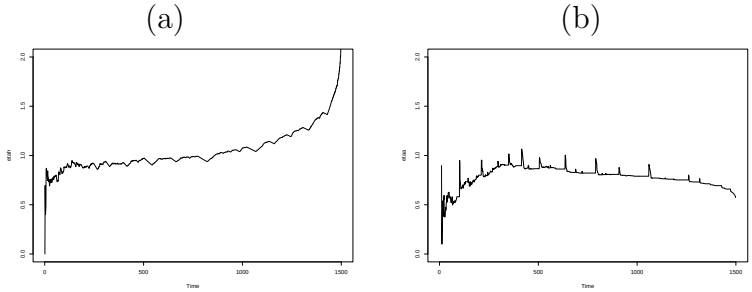


Figure 11: Estimators for η (a) $(k, \widehat{\eta}_{H,k})$; (b)$(k, \widehat{\eta}_{B,k})$.

The empirical probability is 0.006.

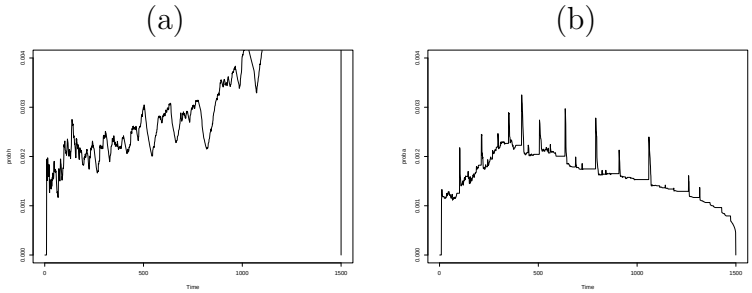


Figure 12: Estimators for $\mathbb{P}(X_1 > 200000, X_2 > 100000)$ (a) $(k, \widehat{p}_{H,k})$; (b) $(k, \widehat{p}_{B,k})$ with the empirical probability added with a horizontal line.

5.1.1 Data set HANES

We consider the variables X_1=Standing Height (cm) and X_2 Weight (kg) for females from the National Health and Nutrition Examination Survey (NHANES) 2005-2006 (National Center for Health Statistics, 2007). Data of females of age 18-64 years only have been retained; this leads to a sample size of 2237. It might be interesting to estimate the proportion of 'tall' and 'heavy' females and estimate $\mathbb{P}(X_1 > 175, X_2 > 125)$.

Again first the bivariate data are transformed into Fréchet marginals (Z_1, Z_2),using the empirical distribution function. To estimate $\mathbb{P}(X_1 > 175, X_2 > 125)$, it is transformed into Fréchet marginals to become $\mathbb{P}(Z_1 > z_1, Z_2 > z_2)$ which can be estimated as before. We choose $w = z_1/(z_1 + z_2) = 0.49$.
The empirical probability is 0.0013 and is added on both graphs below.

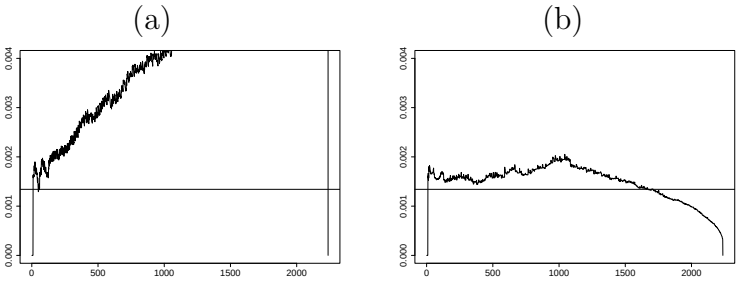


Figure 13: Estimators for $\mathbb{P}(X_1 > 175, X_2 > 125)$ (a) $(k, \widehat{p}_{H,k})$; (b) $(k, \widehat{p}_{B,k})$ with the empirical probability added with a horizontal line.

The estimators for η are given in Figure 14.

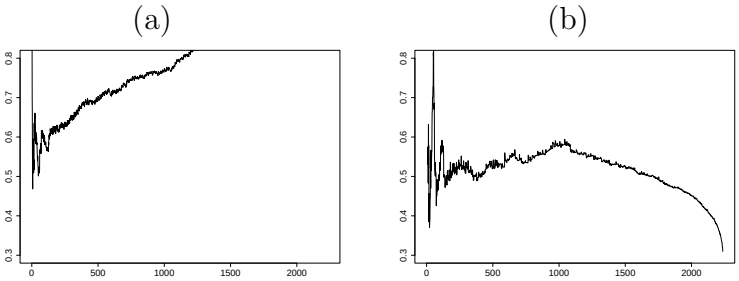


Figure 14: Estimators for η (a) $(k, \widehat{\eta}_{H,k})$; (b)$(k, \widehat{\eta}_{B,k})$.

6 Conclusion

Based on some heuristic arguments, biased reduced estimators η_H and $\widehat{p}_B$ were introduced. Simulations show that the bias reduced estimators seem to work quite well for the coefficient of tail dependence as well as for small probabilities as far as the bias is concerned. Moreover, the estimators do not depend as heavily on the number of extremes taken into account in the

estimation. Also in real life examples the bias reduced estimator seems to provide reasonable results.

References

[1] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2005). Statistics of Extremes: Theory and Applications. *Wiley*, 490p.

[2] Fraga Alves, M.I., Gomes, M.I., and de Haan, L. (2003) A new class of semi- parametric estimators of the second order parameter. *Portugaliae Mathematica* **60**, 193-214.

[3] 1998Frees98 Frees, E.W., and Valdez E.A. (1998) Understanding relationships using copulas. *North American Actuarial Journal* **2**, 1-15.

[4] 1999Klugman Klugman S.A., and Parsa R. (1999) Fitting bivariate loss distributions with copulas. *Insurance: Mathematics and Economics* **24**, 139-148.

[5] Ledford, A.W. and Tawn J.A. (1997) Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society B*, **59**, 475-499.

[6] National Center for Health Statistics (2007). National Health and Nutrition Examination Survey 2005-2006. Body measurement data, http://www.cdc.gov/nchs/about/major/nhanes/nhanes2005-2006/exam05$_0$6.htm.

A Simple Estimation Procedure for Categorical Data Analysis

Nico Crowther
Department of Statistics, University of Pretoria

1 MLE Estimation Under Constraints

Consider the random vector $\boldsymbol{x} : p \times 1$ which has a normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ may be singular. For any constant matrix $\boldsymbol{G}$, the conditional distribution of $\boldsymbol{x}$, given that $\boldsymbol{Gx} = \boldsymbol{c}$, is normal with

$$\begin{aligned} E(\boldsymbol{x}|\boldsymbol{Gx} = \boldsymbol{c}) &= \boldsymbol{\mu} + (\boldsymbol{G\Sigma})'(\boldsymbol{G\Sigma G}')^*(\boldsymbol{c} - \boldsymbol{G\mu}) \\ Cov(\boldsymbol{x}|\boldsymbol{Gx} = \boldsymbol{c}) &= \boldsymbol{\Sigma} - (\boldsymbol{G\Sigma})'(\boldsymbol{G\Sigma G}')^*\boldsymbol{G\Sigma} = \boldsymbol{\Sigma}_C \end{aligned}$$

where $(\boldsymbol{G\Sigma G}')^*$ denotes any generalized inverse of $(\boldsymbol{G\Sigma G}')$ and

$$rank(\boldsymbol{\Sigma}_C) = rank(\boldsymbol{\Sigma}) - rank(\boldsymbol{G\Sigma}) \quad .$$

If $\boldsymbol{\Sigma}$ is non-singular and known, the maximum of the density function

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}$$

of $\boldsymbol{x}$, given that $\boldsymbol{Gx} = \boldsymbol{c}$, is obtained at the mean of the conditional distribution of $\boldsymbol{x}$, given that $\boldsymbol{Gx} = \boldsymbol{c}$, i.e. at

$$\boldsymbol{x} = \boldsymbol{\mu} + (\boldsymbol{G\Sigma})'(\boldsymbol{G\Sigma G}')^*(\boldsymbol{c} - \boldsymbol{G\mu}) \quad .$$

Since the roles of $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are symmetric, the maximum of the likelihood function

$$L(\boldsymbol{\mu}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}$$

under the restriction that $\boldsymbol{G\mu} = \boldsymbol{c}$, is obtained at

$$\widehat{\boldsymbol{\mu}}_c = \boldsymbol{x} + (\boldsymbol{G\Sigma})'(\boldsymbol{G\Sigma G}')^*(\boldsymbol{c} - \boldsymbol{Gx}) \quad . \qquad (1)$$

If $\boldsymbol{\Sigma}$ is known, the expression above gives an explicit expression for $\widehat{\boldsymbol{\mu}}$, the MLE for $\boldsymbol{\mu}$ under the restriction that $\boldsymbol{G\mu} = \boldsymbol{c}$. If $\boldsymbol{\Sigma}$ is a function of $\boldsymbol{\mu}$, the maximum likelihood estimator of $\boldsymbol{\mu}$ can be obtained iteratively.

Although it is assumed in the discussion above that $\boldsymbol{x}$ has a normal distribution, the expression (1) for obtaining an MLE can be adjusted to hold for any member of the exponential class. (See Matthews and Crowther (1995).)

The basic idea of obtaining an MLE as presented in (1), may also be extended to obtain the MLE for the frequency vector $\boldsymbol{F}$ under the restriction that $g(\boldsymbol{F}) = \boldsymbol{0}$. If $\boldsymbol{f}$ is an unrestricted MLE of $\boldsymbol{F}$ with covariance matrix $\boldsymbol{\Sigma}$ and $\frac{\partial \boldsymbol{g}(\boldsymbol{F})}{\partial \boldsymbol{F}} = \boldsymbol{G}$, the MLE of $\boldsymbol{F}$ under the restriction that $g(\boldsymbol{F}) = \boldsymbol{0}$, may be obtained iteratively from

$$\widehat{\boldsymbol{F}} = \boldsymbol{f} - (\boldsymbol{G\Sigma})'(\boldsymbol{G\Sigma G}')^* g(\boldsymbol{f}) \quad . \qquad (2)$$

If $g(\boldsymbol{F})$ is linear in $ln(\boldsymbol{F})$, say $g(\boldsymbol{F}) = \boldsymbol{A}'ln(\boldsymbol{F})$, and if the columns of $\boldsymbol{A}$ are orthogonal to the vector of ones, then under standard sampling procedures like Poisson, multinomial or product multinomial, the procedure in (2) reduces to (with $\boldsymbol{D}_F = diag(\boldsymbol{F})$):

$$\widehat{\boldsymbol{F}} = \boldsymbol{f} - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{D}_F^{-1}\boldsymbol{A})^*\boldsymbol{A}'ln(\boldsymbol{f}) \quad . \qquad (3)$$

This expression produces the exact MLE. It also illustrates the fact that in this case the different sampling procedures produce the same MLE's.

The Wald statistic for estimating the goodness of fit is:

$$W = g(\boldsymbol{f})'(\boldsymbol{G\Sigma G}')^* g(\boldsymbol{f}) \quad .$$

Example 1 Let $\boldsymbol{f}' = \{65, 15, 10\}$ be an observation from a trinomial distribution. Suppose that the model to be fitted is

$$F_1 F_3 - F_2^2 = 0 \quad \text{or} \quad log(F_1) - 2log(F_2) + log(F_3) = 0$$

In this case:

$$\boldsymbol{A}' = \{1, -2, 1\} \qquad \text{and} \qquad \boldsymbol{A}'\boldsymbol{D}_F^{-1}\boldsymbol{A} = \frac{1}{F_1} + \frac{4}{F_2} + \frac{1}{F_1}$$

Hence

$$\widehat{\boldsymbol{F}} = \boldsymbol{f} - e \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

with $e = \dfrac{log(F_1) - 2log(F_2) + log(F_3)}{\frac{1}{F_1} + \frac{4}{F_2} + \frac{1}{F_3}}$. The successive iterations are:

e	$\widehat{\boldsymbol{F}}$	e	$\widehat{\boldsymbol{F}}$	e	$\widehat{\boldsymbol{F}}$	e	$\widehat{\boldsymbol{F}}$
2.776779	62.223221	0.1774744	62.045747	−0.000046	62.045793	$-6.4E-12$	62.045793
	20.553558		20.908507		20.908414		20.908414
	7.223221		7.0457466		7.0457928		7.0457928

The Wald statistic is $W = 2.95$ and a χ_1^2 distribution suggests a bad fit.

2 Constructing Value Models

In many instances a frequency table represents an underlying value or assessment as in the following example from the book of Agresti (1990).

Example 2

Preference for Black Olives, by Urbanization and Location*

		Preference						Means	
Scale		1	2.5	4.5	6	7	8.5		
Urbanization	**Location**	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{C}$	$\boldsymbol{D}$	$\boldsymbol{E}$	$\boldsymbol{F}$	**Sample**	**Predicted**
Urban	$\boldsymbol{MW}$	20	15	12	17	16	28	5.22	4.87
	$\boldsymbol{NE}$	18	17	18	18	06	25	4.94	5.00
	$\boldsymbol{SW}$	12	09	23	21	19	30	5.72	5.92
Rural	$\boldsymbol{MW}$	30	22	21	17	8	12	4.00	4.28
	$\boldsymbol{NE}$	23	18	20	18	10	15	4.46	4.41
	$\boldsymbol{SW}$	11	09	26	19	17	24	5.54	5.33

* $\boldsymbol{A}$ dislike extremely; $\boldsymbol{B}$, dislike very much or moderately; $\boldsymbol{C}$ dislike slightly or neither like or dislike; $\boldsymbol{D}$, like slightly; $\boldsymbol{E}$, like moderately; $\boldsymbol{F}$, like very much or like extremely

The sample consists of six independent samples from the different locations.

Using a six-point ordinal scale with values representing preference as indicated, subjects reported their preference for black olives. The construction of the value scale is of fundamental importance since the objectiveness of the statistical results depends directly on it.

The frequencies and model to be fitted are as follows:

$$\begin{aligned} \boldsymbol{F}' &= (\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4, \boldsymbol{F}_5, \boldsymbol{F}_6) \\ &= (n_1\boldsymbol{p}_1, n_2\boldsymbol{p}_2, n_3\boldsymbol{p}_3, n_4\boldsymbol{p}_4, n_5\boldsymbol{p}_5, n_6\boldsymbol{p}_6) \end{aligned}$$

$$\boldsymbol{\nu}' = (1, 2.5, 4.5, 6, 7, 8.5)\,.$$

and

$$\begin{pmatrix} \boldsymbol{\nu}'\boldsymbol{p}_1 \\ \boldsymbol{\nu}'\boldsymbol{p}_2 \\ \boldsymbol{\nu}'\boldsymbol{p}_3 \\ \boldsymbol{\nu}'\boldsymbol{p}_4 \\ \boldsymbol{\nu}'\boldsymbol{p}_5 \\ \boldsymbol{\nu}'\boldsymbol{p}_6 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \boldsymbol{A}\boldsymbol{\lambda}$$

Using (2) it follows that:

$$\begin{array}{lcl} \boldsymbol{\Sigma}_i & = & \left(\boldsymbol{D}_{p_i} - \boldsymbol{p}_i\boldsymbol{p}_i'\right)/n_i \\ \boldsymbol{\Sigma} & = & block(\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_6) \\ \boldsymbol{D} & = & \boldsymbol{I}(nr) \otimes \boldsymbol{\nu}' \\ \boldsymbol{G} & = & [\boldsymbol{I}(nr) - \boldsymbol{A} * inv(\boldsymbol{A}' * \boldsymbol{A}) * \boldsymbol{A}'] * \boldsymbol{D} \\ \boldsymbol{g} & = & \boldsymbol{G} * \boldsymbol{p} \end{array}$$

A SAS program for computing the MLE's is given in the appendix.

The MLE's of the parameters are given below.

	Iteration Number					Location	Sample	Pred
	1	2	3	4	5	$\boldsymbol{MW}$	5.22	4.87
μ	4.96807	4.96484	4.96497	4.96495	4.96496	$\boldsymbol{NE}$	4.94	5.00
α_1	0.29654	0.29340	0.29341	0.29337	0.29337	$\boldsymbol{SW}$	5.72	5.91
α_2	-0.29654	-0.29340	-0.29341	-0.29337	-0.29337	$\boldsymbol{MW}$	4.00	4.28
β_1	-0.39333	-0.39037	-0.39037	-0.39029	-0.39029	$\boldsymbol{NE}$	4.46	4.41
β_2	-0.26618	-0.26312	-0.26325	-0.26324	-0.26324	$\boldsymbol{SW}$	5.54	5.33
β_2	0.65951	0.65349	0.65362	0.65352	0.65353			

Agresti (1990) obtained weighted least squares estimates for the parameters, which of coarse equals the first column in the table above.

The Wald statistic for the fit of the model is 4.71 with two degrees of freedom.

The Wald statistic for equating $\boldsymbol{MW}$ and $\boldsymbol{NE}$ is 0.126 with one degree of freedom.

The MLE's of the parameters, equating $\boldsymbol{MW}$ and $\boldsymbol{NE}$ are as follows:

	Iteration Number					Location	Sample	Pred
	1	2	3	4	5	$\boldsymbol{MW}$	5.22	4.93
μ	4.96700	4.96451	4.96518	4.96518	4.96519	$\boldsymbol{NE}$	4.94	4.93
α_1	0.29745	0.29422	0.29413	0.29408	0.29408	$\boldsymbol{SW}$	5.72	5.91
α_2	-0.29745	-0.29422	-0.29413	-0.29408	-0.29408	$\boldsymbol{MW}$	4.00	4.34
β_1	-0.33028	-0.32686	-0.32667	-0.32661	-0.32661	$\boldsymbol{NE}$	4.46	4.34
β_2	-0.33028	-0.32686	-0.32667	-0.32661	-0.32661	$\boldsymbol{SW}$	5.54	5.32
β_2	0.66057	0.65372	0.65334	0.65322	0.65322			

The Wald statistic for the fit of this model is 4.95 with three degrees of freedom.

The Wald statistic for equating $\boldsymbol{MW}$, $\boldsymbol{NE}$ and $\boldsymbol{SW}$ is 22.6 with two degrees of freedom.

3 A Typical Value Model for Insurance Policies

Example 3

This example illustrates a typical value model for the classification of insurance policies. A value score can be constructed for each policy, taking factors like the status of a policy, losses incurred due to cancellations, the profile of the client and other typical factors into account. In the case of a large data set, the only requirement is that value of the policies should be measured on an ordinal scale. The value score of a policy can then be reassigned on a proper scale, say from 0 to 100.

The parameters in the value model may be main effects or interaction effects. In this example only main effects are taken into account. It should be noted that the effects over the categories of every explanatory variable add up to zero. The overall mean effect, using a 0-100 scale, should be near 50. In the case of an equal number of observations in each category of an explanatory variable, it should add up to zero. The predictor "VALUE AREA" is created by grouping postal codes of policies with relative similar values together.

The estimated value of every policy may now be calculated, using the effects in the table below.

LIFE INSURANCE: One-Year Value Model

Category	n	Effect
Mean	376248	51.76
AGE and GENDER		
-25,female	47032	-6.17
-25,male	41958	-4.82
25-30,female	36103	-2.30
25-30,male	31757	-1.17
30-35,female	33088	0.09
30-35,male	27490	0.32
35-45,female	52142	1.01
35-45,male	35939	2.89
45+,female	42603	4.16
45+,male	28136	6.00
POLICY		
Product A:Phone	31680	-0.60
Product A:Web	29213	-1.74
Product B:Web	104000	0.06
Product C	211355	2.28
BANK ACCOUNT		
BANK A:Cheque	18824	3.44
BANK A:Savings	112379	-2.60
BANK B:Savings	27161	-8.41
BANK C:Cheque	52057	5.96
BANK C:Savings	32892	4.96
BANK D:Savings	26563	-11.05
BANK E:Cheque	9963	5.29
BANK E:Savings	96409	2.41

Category	n	Effect
DEBIT DAY		
01	132248	-2.38
02-14	61154	-6.33
15	31739	0.83
16-24	19376	-1.55
25	34837	2.29
26	10730	6.55
27-29	21213	3.73
last day	64951	-3.14
VALUE AREA		
A	38524	7.64
B	70990	3.82
C	154431	0.00
D	73464	-3.82
E	38839	-7.64

4 References

Matthews, G.B. and Crowther, N.A.S. (1995) A maximum likelihood estimation procedure when modelling in terms of constraints. S.A Statist. J., 29, 29-51.

Agresti, A. (1990) *Categorical Data Analysis.* John Wiley & Sons.

5 Appendix

SAS Program for Example 2

```
proc iml;options pagesize=100;
fmat={20 15 12 17 16 28,
      18 17 18 18 06 25,
      12 09 23 21 19 30,
      30 22 21 17  8 12,
      23 18 20 18 10 15,
      11 09 26 19 17 24};
n=fmat[,+];nr=nrow(fmat);
pmat=diag(1/n)*fmat;
p=colvec(pmat);
nu={1,2.5,4.5,6,7,8.5};
A={1  1  1  0,
   1  1  0  1,
   1  1 -1 -1,
   1 -1  1  0,
   1 -1  0  1,
   1 -1 -1 -1};
D=I(nr)@nu';
G1=(I(nr)-A*inv(A'*A)*A')*D;
L=inv(A'*A)*A';
LL=L[1:2,]//(-L[2,])//L[3:4,]//(-L[3:4,])[+,];
*G2=(L[3,]-L[4,])*D;  *For equating NE and SW;
G2=((L[3,]//L[4,]))*D;  *For equating NE,SW and MW;
*G=G1;
G=G1//G2;

diff=1;
i=0;p0=p;
do while (diff>1e-8);
i=i+1;
pv=p;
pmat=shape(p,nr);
sig=(diag(pmat[1,])-pmat[1,]'*pmat[1,])/n[1];
do j=2 to nr;
sigt=(diag(pmat[j,])-pmat[j,]'*pmat[j,])/n[j];
sig=block(sig,sigt);
end;
if i=1 then sig1=sig;

gg=G*p0;
gg2=G2*p0;
p=p0-(G*sig)'*ginv(G*sig*G')*gg;
par=LL*D*p;
print i  par[format=7.5] p;
diff=sqrt((p-pv)'*(p-pv));
end;

Wald=gg'*ginv(G*sig*G')*gg;print Wald;
Wald2=gg2'*ginv(G2*sig*G2')*gg2;print Wald2;
Sample=D*p0;Pred=D*p;print sample[format=4.2] pred[format=4.2];
```

STATISTICAL CHARACTERIZATION OF QT PROLONGATION

Robert Schall[1,2] **and Arne Ring**[3,4]

[1] *Department of Mathematical Statistics and Actuarial Science, University of the Free State, 9300 Bloemfontein, South Africa*

[2] *Quintiles Biostatistics, PO Box 12603, Bloemfontein 9324, South Africa*

[3] *Clinical Biostatistics, Boehringer Ingelheim Pharma, 88400 Biberach an der Riss, Germany*

[4] *Institute for Biometry, University of Ulm, 89075 Ulm, Germany*

Key Words: Bazett; ECG; Fridericia; Heart rate correction; Mixed models; QT interval; Thorough QT study; Torsades des pointes.

ABSTRACT

The appropriate assessment of QT prolongation remains controversial. We suggest that, before the relative merits of various methods can be evaluated, we must state what we assume an assessment of QT prolongation should be about. As a general framework for the assessment of QT prolongation we propose that an assessment of "absolute" or "uncorrected" QT prolongation is properly carried out through a between-treatment (active versus placebo) comparison of the marginal distributions of QT data; an assessment of "heart rate corrected" QT prolongation is carried out through a between-treatment comparison of the conditional distributions of QT data (conditional on RR interval or heart rate). Under this general framework, conditional QT prolongation is, in general, a function of RR interval, and we discuss three possible summary characteristics for that function. We show how current procedures for the assessment of QT prolongation relate to the general approach (that is, to between-treatment contrasts of the marginal and conditional expectation of the QT interval), and to each other. It transpires that only the so-called "one-step procedure" can provide a complete characterization of conditional QT prolongation. We show that the "two-step procedure" with data driven correction provides an unbiased estimate of *expected conditional QT prolongation*, which may, from a clinical point of view, be a more satisfactory characteristic than the conventional characteristic, QT prolongation at the reference RR interval. We strongly suggest that two-step procedures with fixed correction be abandoned in the analysis of thorough QT/QTc studies: fixed correction is either redundant (when a drug has no effect on RR average interval), or systematically biased (when a drug does affect average RR interval).

1. INTRODUCTION

1.1. QT Prolongation

An undesirable property of some non-anti-arrhythmic drugs is their ability to delay cardiac repolarization, an effect that can be measured as prolongation of the QT interval on the surface electrocardiogram (ECG) (ICH 2005). A delay in cardiac repolarization is associated with the development of arrhythmia, in particular Torsades de Pointes, which may be life-threatening. Thus prolongation of the QT interval is used as a biomarker for cardiac safety. In recent years, regulatory authorities require that every new drug undergo clinical electrographic evaluation. Typically QT data is collected early in the drug's development, to be followed by a so-called "thorough QT/QTc study" in healthy subjects (ICH 2005). If a drug's potential for QT prolongation cannot be ruled out based on the results of those studies, an expanded ECG safety evaluation during later stages of drug development must be conducted, including clinical trials in the target population. Alternatively, the drug's development might be terminated. Either way, the assessment of QT prolongation is a crucial part of drug development and in recent years has generated intensive discussion in the medical and statistical literature.

1.2. Effect of RR Interval

Because the QT interval is associated with the RR interval (which is the inverse of the heart rate), the potential of a drug to prolong the QT interval could be "masked" if the drug decreases the RR interval. Conversely, a drug that does not prolong the QT interval may appear to do so if the drug increases the RR interval. Therefore, when a drug affects the RR interval, the drug's effect on the QT interval must be distinguished from changes in QT interval due to changes in RR interval (Li et al, 2004). For this reason, QT interval data is routinely "corrected" for RR interval using various correction formulae or procedures, to obtain a value known as the QTc interval which is hoped to be independent of RR interval, or at least less dependent of RR interval than the QT interval itself (ICH 2005). However, as the ICH (2005) guidance

document states, it is not yet clear whether cardiac arrhythmia is more closely related to an increase in the absolute (ie. uncorrected) QT interval or to an increase in QTc. As long is this clinical question is open, both QT and QTc data have to be studied.

1.3. Current Approaches to Assessment of QT Prolongation

Historically, the analysis of QT data proceeds in two steps (Li et al, 2004): In the first step, the QT interval data is corrected for the effect of heart rate, using one of the following two correction procedures: either (i) "Fixed" correction where one of a number of published correction formulae is used to calculate QTc as a function of QT and HR; or (ii) "Off-drug data driven" correction where the slope estimate from a regression of placebo (off-drug) QT data against heart rate is used to calculate QTc for both active and placebo treatment data. Thereafter, in the second step, the QTc data of active and placebo treatment are compared statistically.

More recently, Li et al (2004) and other authors have pointed out that, from a statistician's perspective, a so-called "one-step" statistical analysis of both active treatment and placebo QT data, using mixed model analysis of covariance (where RR is fitted as a covariate) would be preferable to the two-step approach. Nevertheless, many clinical pharmacologists and other medical researchers seem to be comfortable with the two-step analysis and the ICH (2005) guidance as well as most analyses of QT/QTc trials in practice adhere to the two-step approach. The situation persists despite published work pointing out serious shortcomings of the various two-step approaches (Dmitrienko and Smith, 2003; Shah and Hajian, 2003; Li et al, 2004; Wang, Pan and Balch, 2008).

1.4. Outline of the Rest of the Paper

The appropriate assessment of QT prolongation evidently remains controversial. We suggest that, before the relative merits of the various methods for assessment of QT prolongation can be evaluated, a basis for discussion is needed. At a minimum, we must explicitly state what we assume an assessment of QT prolongation is (or should be) about. In Section 2, therefore, we propose a general and basic principle for the

assessment of QT prolongation: We suggest that an assessment of uncorrected or "absolute" QT prolongation is properly carried out through a between-treatment (active versus placebo) comparison of the marginal distributions of QT data; an assessment of "heart rate corrected" QT prolongation is carried out through a between-treatment comparison of the conditional distributions of QT data (conditional on RR interval). Current procedures (analysis of central tendency and categorical analysis of QT prolongation (ICH 2005, Section 3.2)) emerge as special cases.

Both the relevant literature and current practice focus on the case where the conditional expectation of QT data is written as a linear or log-linear function of the RR interval. In Section 3 we adopt this approach and argue that, in general, the linear or log-linear model should allow for unequal slopes for the regression of QT against RR interval. Thus conditional QT prolongation is, in general, a function of RR interval. We discuss three possible summary characteristics for that function, and show how the various characteristics for assessment of both marginal and conditional QT prolongation can be given a geometric interpretation.

In Section 4 we show how current procedures for assessment of QT prolongation relate to the general approach (that is, to between-treatment contrasts of the marginal and conditional expectation of the QT interval), and to each other. It transpires that, in general, only the one-step procedure provides a complete characterization of conditional QT prolongation. We show that the two-step procedure with data driven correction provides an unbiased estimate of *expected conditional QT prolongation*, which may, from a clinical point of view, be a more satisfactory characteristic than the conventional characteristic, QT prolongation at the reference RR interval. The two-step procedure with fixed correction is either redundant (when a drug has no effect on RR average interval), or potentially biased (when a drug does affect average RR interval).

In Section 5 we sketch a procedure and statistical decision rule for assessment of QT prolongation. When QT data is analyzed on the logarithmic scale, we propose to formulate the statistical decision rule on the ratio scale. Section 6 is devoted to an

example of application, and Section 7 to the discussion.

In summary, our paper focuses on the *parametric* characterization of QT prolongation. We aim to answer the following question: which parameter contrast is statistical inference about under various current QT assessment procedures (one-step procedure, two-step procedure with data driven correction, two-step procedure with fixed correction, analysis of uncorrected data). We also make a suggestion regarding which parameter contrast statistical inference should be about in our view. Details of the estimation or testing of the parameter contrasts in question are not of primary interest in the present paper.

2. MARGINAL VERSUS CONDITIONAL QT PROLONGATION

Let QT denote a QT interval measurement and RR the corresponding RR interval. We define $y = f(QT)$ and $x = f(RR)$, where $f(\cdot)$ is a suitable monotonic transformation. Usually, $f(\cdot)$ is either the identity function or the natural logarithm (Malik et al. 2002; Ring, 2009). The bivariate random variables (y_A, x_A) and (y_P, x_P) denote the paired (possibly transformed) QT and RR interval measurements under active (A) and placebo (P) treatment, respectively. Furthermore, F_A and F_P denote the marginal distributions of y_A and y_P, $F_A|x$ and $F_P|x$ the conditional distributions of y_A and y_P (conditional on a given value of x), and G_A and G_P the marginal distributions of x_A and x_P.

We define the assessment of "uncorrected" or "absolute" QT prolongation as a between-treatment (active versus placebo) comparison of the marginal distributions F_A and F_P; in the following, we will refer to such a comparison as an assessment of *marginal* QT prolongation. Similarly, we define the assessment of "heart rate corrected" QT prolongation as a between-treatment comparison of the conditional distributions $F_A|x$ and $F_P|x$; we will refer to such a comparison as an assessment of *conditional* QT prolongation.

While in general an assessment of marginal and conditional QT prolongation involves a comparison of marginal and conditional distributions of QT measurements,

in practice such a comparison will focus on specific parameters or functions of those distributions. Below, we present two such types of characteristics.

2.1. Characteristics for Marginal QT Prolongation

An assessment of marginal QT prolongation may focus on the means $\mathrm{E}(y_A) = \mathrm{E}_{F_A}(y_A) = \mu_A$ and $\mathrm{E}(y_P) = \mathrm{E}_{F_P}(y_P) = \mu_P$ of the relevant marginal distributions F_A and F_P; in the terminology of the relevant ICH guidance (ICH 2005, Section 3.2.1) a comparison of means constitutes an analysis of the "central tendency" of QT measurements. Then the between-treatment contrast of marginal means,

$$\gamma_m = \mu_A - \mu_P \tag{1}$$

can be viewed as a "moment-based" characteristic for marginal QT prolongation, to use terminology from area of bioequivalence assessment (Schall and Luus, 1993). (The subscript "m" in γ_m stands for "marginal".)

Of course the difference in means is not the only criterion that can be used to compare the distributions F_A and F_P in a clinically meaningful way. For example, so-called "categorical" analyses (ICH 2005, Section 3.2.2) involve estimation of excess probabilities of the form $\mathrm{Prob}_{F_A}(y_A > c)$ and $\mathrm{Prob}_{F_P}(y_P > c)$, where c represents a clinically relevant cut-off value for y, such as $c = f(450\mathrm{ms})$, $c = f(480\mathrm{ms})$ or $c = f(500\mathrm{ms})$. A "probability-based" characteristic for marginal QT prolongation could then be defined as the risk ratio

$$\rho_{m,d} = \frac{\mathrm{Prob}_{F_A}(y_A > d)}{\mathrm{Prob}_{F_P}(y_P > d)} \tag{2}$$

Alternatively, the odds ratio or risk difference involving the excess probabilities for active and placebo treatment could be used to characterize relative risk.

2.2. Characteristics for Conditional QT Prolongation

An assessment of conditional QT prolongation may focus on the conditional means $\mathrm{E}(y_A|x) = \mathrm{E}_{F_A|x}(y_A)$ and $\mathrm{E}(y_P|x) = \mathrm{E}_{F_P|x}(y_P)$ of y_A and y_P. Then, a moment-based

characteristic for conditional QT prolongation is the between-treatment contrast of conditional means

$$\gamma_c(x) = \mathrm{E}(y_A|x) - \mathrm{E}(y_P|x) \tag{3}$$

which in general is a function of x. (The subscript "c" in $\gamma_c(x)$ stands for "conditional".) We note that Equation (3) does not necessarily require a linear model for the conditional means. In principle, $\mathrm{E}(y_A|x)$ and $\mathrm{E}(y_P|x)$, and consequently $\gamma(x)$, can be estimated using non-linear or nonparametric regression techniques. However, in the following we concentrate on the case when $\mathrm{E}(y_A|x)$ and $\mathrm{E}(y_P|x)$ can be written as linear functions of x.

Similarly to Equation (2), a probability-based characteristic for conditional QT prolongation can be defined as the risk ratio

$$\rho_{c,d}(x) = \frac{\mathrm{Prob}_{F_A|x}(y_A > d)}{\mathrm{Prob}_{F_P|x}(y_P > d)} = \frac{\mathrm{Prob}_{F_A}(y_A > d \mid x)}{\mathrm{Prob}_{F_P}(y_P > d \mid x)} \tag{4}$$

3. CHARACTERIZATION OF QT PROLONGATION UNDER LINEAR/LOG-LINEAR MODEL

3.1. Linear/Log-linear Model for QT Data

In the spirit of Dmitrienko and Smith (2003), Shah and Hajian (2003) and Li et al (2004) we postulate the following linear model for the conditional expectations $\mathrm{E}(y_A|x)$ and $\mathrm{E}(y_P|x)$ as a function of x:

$$\mathrm{E}(y_i|x) = \alpha_i + \beta_i \cdot (x - x_r); \qquad i = A, P \tag{5}$$

Here $x_r = f(RR_r) = f(1000)$, where $RR_r = 1000$ ms is the fixed value of the so-called "reference" RR interval corresponding to a heart rate of 60 beats/min. Note that in Model (5) we allow, in general, for different (non-parallel) slopes β_A and β_P for active and placebo treatment, respectively. When $f(\cdot)$ is the identity function, Model (5) is linear; when $f(\cdot)$ is the logarithm, Model (5) is log-linear.

In terms of the parameters of Model (5), and of the marginal means $\mathrm{E}_{G_A}(x_A) = \mathrm{E}(x_A) = \nu_A$ and $\mathrm{E}_{G_P}(x_P) = \mathrm{E}(x_P) = \nu_P$ of x_A and x_P, the unconditional expectations

of y_A and y_P are given by

$$\mathrm{E}(y_i) = \mu_i = \alpha_i + \beta_i \cdot (\nu_i - x_r); \qquad i = A, P \tag{6}$$

We can now express the moment-based characteristics for marginal and conditional QT prolongation – Equations (1) and (3), respectively – in terms of the parameters of Model (5) and of the expected values of x_A and x_P.

To illustrate the various characteristics for QT prolongation, in Figure 1 we present a schematic plot of the regression lines specified in Model (5), namely of $\mathrm{E}(y_A|x)$ and $\mathrm{E}(y_P|x)$. For simplicity, but without loss of generality we assume that (5) represents a log-linear model with $x_r = \log(1[\mathrm{s}]) = 0$.

3.2. Marginal QT prolongation

Although not necessary for the assessment of marginal QT prolongation, it will be instructive to express γ_m in terms of the parameters of Model (5), using Equation (6):

$$\gamma_m = \mu_A - \mu_P = [\alpha_A + \beta_A \cdot (\nu_A - x_r)] - [\alpha_P + \beta_P \cdot (\nu_P - x_r)] \tag{7}$$

Thus γ_m, the treatment contrast for assessment of marginal QT prolongation, can be written as $\gamma_m = \mathrm{E}(y_A|\nu_A) - \mathrm{E}(y_P|\nu_P)$.

We note that the contrast γ_m can be obtained geometrically as follows: project the points $\mathrm{E}(y_A|\nu_A)$ and $\mathrm{E}(y_P|\nu_P)$ *parallel* with the x-axis (that is, using a *slope of zero*) onto the y-axis; γ_m is then given by the difference of the projections of $\mathrm{E}(y_A|\nu_A)$ and $\mathrm{E}(y_P|\nu_P)$ onto the y-axis (Figure 1). Below we will show how the various characteristics for conditional QT prolongation can also be obtained geometrically, namely as the difference of projections of $\mathrm{E}(y_A|\nu_A)$ and $\mathrm{E}(y_P|\nu_P)$ onto the y-axis using particular sets of regression slopes (Table 1).

3.3. Conditional QT prolongation

Using Equation (5), we can write the between-treatment contrast of conditional means, $\gamma_c(x)$ in (3), as

$$\gamma_c(x) = \mathrm{E}(y_A|x) - \mathrm{E}(y_P|x) = \alpha_A - \alpha_P + (\beta_A - \beta_P) \cdot (x - x_r) \tag{8}$$

When, in Model (5), we are willing to assume that the slopes are equal ($\beta_A = \beta_P$), $\gamma_c(x)$ simplifies to $\gamma_c(x) \equiv \gamma_c \equiv \alpha_A - \alpha_P$, and conditional QT prolongation is characterized globally (independent of RR interval) by the difference in intercepts $\alpha_A - \alpha_P$. However, in general the assumption of equal slopes cannot be made so that the extent of conditional QT prolongation, $\gamma_c(x)$, depends on $x = f(RR)$. For use in a statistical decision rule, we might consider the following three summary characteristics for conditional QT prolongation: (i) conditional QT prolongation $\gamma_c(x)$ evaluated at some reference RR interval; (ii) the expected value of conditional QT prolongation $\gamma_c(x)$; (iii) the maximum conditional QT prolongation $\gamma_c(x)$ over a "reference range" of RR interval values, that is, the maximum of $\gamma_c(x)$ over a clinically relevant range $[x_0, x_1]$ of values of x.

(i) Conditional QT prolongation at some reference RR interval

Characteristic $\gamma_c(x)$ evaluated at the conventional reference RR interval is given by

$$\gamma_c(x_r) = \alpha_A - \alpha_P \tag{9}$$

Characteristic (9) is commonly viewed almost as the "definition" of (conditional) QT prolongation. However, in the case of unequal slopes under Model (5), $\gamma_c(x_r)$ characterizes conditional QT prolongation only for one particular RR interval, namely for the conventional reference RR interval $RR_r = 1$ s, and one may suspect that $RR_r = 1$ s was not necessarily chosen for its clinical relevance but simply because it is a "round" number. Therefore, regulatory authorities might consider changing the definition of reference RR interval, for example to a value that better represents the typical RR interval seen in the patient target population.

Geometrically, $\gamma_c(x_r)$ can be obtained as follows: project the points $\mathrm{E}(y_A|\nu_A)$ and $\mathrm{E}(y_P|\nu_P)$ onto the y-axis using the *respective regression slopes* β_A *and* β_P; $\gamma_c(x_r)$ is then given by the difference of these projections, namely as the difference of intercepts $\alpha_A - \alpha_P$ (Figure 1; Table 1).

(ii) Expected value of conditional QT prolongation

The expected value of $\gamma_c(x)$, taking the expectation with respect to the distribution G_A of $x = f(RR)$ under active treatment, is given by

$$\mathrm{E}_{G_A}\{\gamma_c(x)\} = \alpha_A - \alpha_P + (\beta_A - \beta_P) \cdot (\nu_A - x_r) = \gamma_c(\nu_A) \tag{10}$$

Characteristic (10) can be interpreted as contrast $\gamma_c(x)$ evaluated at the expected value ν_A of x under active treatment.

Geometrically, $\gamma_c(\nu_A)$ can be obtained as follows: project *both,* $\mathrm{E}(y_A|\nu_A)$ *and* $\mathrm{E}(y_P|\nu_P)$, onto the y-axis using the *regression slope* β_P *for placebo*; $\gamma_c(\nu_A)$ is then given by the difference of the projections (Figure 1; Table 1).

(iii) Maximum conditional QT prolongation over RR reference range

If $x \in [x_0, x_1]$ is a clinically relevant range of values of x, which we refer to as the "reference range" of RR interval (or heart rate) values, then the maximum conditional QT prolongation over that range is given by

$$\gamma_{c,\max} = \max\{\gamma_c(x) \mid x_0 \leq x \leq x_1\} = \max\{\gamma(x_0), \gamma(x_1)\} \tag{11}$$

One could view characteristic (11) as a generalization of characteristic (9), in the sense that instead of considering conditional QT prolongation only at one particular value of RR interval, (maximum) conditional QT prolongation over a range of values is considered. Thorough QT/QTc studies are usually conducted in healthy young subjects, whose average heart rate may well be around 60 (although heart rates under 50 beats/min are often observed). However, the distribution of heart rates of a potential target patient population is likely to be shifted to the right. By specifying $[x_0, x_1]$ so as to represent a range of heart rates from 50 to 70 beats/min, say, (11) would also characterize potential QT prolongation at higher heart rates.

4. CONVENTIONAL PROCEDURES IN THE LIGHT OF THE GENERAL APPROACH

4.1. One-Step Procedure

In the one-step procedure, Model (5) is fitted simultaneously to both active and placebo data, usually through a linear or log-linear mixed model. From the fitted model, unbiased and efficient point and interval estimates for characteristic $\gamma_c(x)$ in (8) can be obtained for any x, that is, over any range of RR interval values of interest (not only for x_r, the reference RR interval). In that sense, the one-step procedure provides a "complete" characterization of conditional QT prolongation. Furthermore, the one-step procedure provides a sound statistical setting for testing specific hypotheses of interest, such as the hypothesis of equal slopes β_A and β_P in Model (5); as we shall see, the question of equal slopes β_A and β_P is highly relevant, both statistically and clinically, to the problem of heart rate correction.

In a conventional application of the one-step procedure, statistical inference is about the parameter contrast $\gamma_c(x_r) = \alpha_A - \alpha_P$, which is summary characteristic (i) of $\gamma_c(x)$ from the general approach: *Conditional QT prolongation at some reference RR interval*. As noted before, however, in the general case of unequal slopes $\beta_A \neq \beta_B$ in Model (5), $\gamma_c(x_r)$ is not necessarily the most appropriate characteristic for conditional QT prolongation and the one-step procedure can be used to characterize conditional QT prolongation more completely.

4.2. Two-Step Procedure with Data Driven Correction

The so-called two-step procedure proceeds as follows: In Step 1, a model of type (5) is fitted to the placebo (or drug-free baseline) data only, namely $\mathrm{E}(y_P|x) = \alpha_P + \beta_P \cdot (x_P - x_r)$. From this model, an estimate $\hat{\beta}_P$ of the slope β_P is obtained. Note that the slope estimate $\hat{\beta}_P$ might have been obtained either as a "pooled" estimate – Model (5) is fitted to the pooled placebo data – or as "individual" estimate – slope estimates are obtained individually for each subject in the placebo group. The slope estimate is then used to correct both the active and placebo data as follows:

$$y_A^c \;=\; y_A - \hat{\beta}_P \cdot (x_A - x_r)$$

$$y_P^c = y_P - \hat{\beta}_P \cdot (x_P - x_r)$$

Thereafter, in Step 2, the heart rate corrected data y_A^c and y_P^c are analyzed. Statistical inference is therefore about the parameter contrast

$$\gamma_2 = \mathrm{E}(y_A^c) - \mathrm{E}(y_P^c) \tag{12}$$

Assuming that the slope estimate $\hat{\beta}_P$ from the placebo data is an unbiased estimate of β_P, we can write $\mathrm{E}(y_P^c)$ and $\mathrm{E}(y_A^c)$ in terms of the parameters of Model (5):

$$\begin{aligned}
\mathrm{E}(y_P^c) &= \mathrm{E}_{G_P}\left\{\mathrm{E}(y_P^c|x)\right\} \\
&= \mathrm{E}_{G_P}\left[\mathrm{E}\left\{y_P - \hat{\beta}_P \cdot (x - x_r) \mid x\right\}\right] \\
&= \mathrm{E}_{G_P}\left\{\mathrm{E}(y_P|x)\right\} - \mathrm{E}_{G_P}\left\{\mathrm{E}(\hat{\beta}_P|x) \cdot (x - x_r)\right\} \\
&= \mathrm{E}_{G_P}\left\{\alpha_P + \beta_P \cdot (x - x_r)\right\} - \mathrm{E}_{G_P}\left\{\beta_P \cdot (x - x_r)\right\} \\
&= \alpha_P + \beta_P \cdot (\nu_P - x_r) - \beta_P \cdot (\nu_P - x_r) \\
&= \alpha_P
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}(y_A^c) &= \mathrm{E}_{G_A}\left\{\mathrm{E}(y_A^c|x)\right\} \\
&= \mathrm{E}_{G_A}\left[\mathrm{E}\left\{y_A - \hat{\beta}_P \cdot (x_A - x_r) \mid x\right\}\right] \\
&= \mathrm{E}_{G_A}\left\{\mathrm{E}(y_A|x)\right\} - \mathrm{E}_{G_A}\left\{\mathrm{E}\left(\hat{\beta}_P|x\right) \cdot (x - x_r)\right\} \\
&= \mathrm{E}_{G_A}\left\{\alpha_A + \beta_A \cdot (x - x_r)\right\} - \mathrm{E}_{G_A}\left\{\beta_P \cdot (x - x_r)\right\} \\
&= \alpha_A + \beta_A \cdot (\nu_A - x_r) - \beta_P \cdot (\nu_A - x_r) \\
&= \alpha_A + (\beta_A - \beta_P) \cdot (\nu_A - x_r)
\end{aligned}$$

The contrast γ_2 in (12) is therefore given by

$$\begin{aligned}
\gamma_2 &= \alpha_A - \alpha_P + (\beta_A - \beta_P) \cdot (\nu_A - x_r) \\
&= \gamma_c(\nu_A)
\end{aligned}$$

Thus the contrast $\gamma_2 = \mathrm{E}(y_A^c) - \mathrm{E}(y_P^c)$ from the two-step procedure with data driven correction is equal to summary characteristic (ii) of $\gamma_c(x)$ from the general approach:

The expected value of conditional QT prolongation (with the expectation taken over the distribution of x under active treatment).

Is the two-step procedure biased?

Li et al (2004) have pointed out that the two-step procedure is biased, and that the procedure implicitly assumes that the slopes β_A and β_P in Model (5) are the same. In the light of the above result, these statements can be put into context: If the slopes β_A and β_P in Model (5) are different, one can indeed view the two-step procedure as providing a biased estimate of the between-treatment contrast $\gamma_c(x_r) = \alpha_A - \alpha_P$, which is the conditional QT prolongation at the reference RR interval; the bias term is $(\beta_A - \beta_P) \cdot (\nu_A - x_r)$. In that sense, Li et al (2004) are correct when claiming that the two-step procedure is biased. However, as shown above, the two-step procedure provides an *unbiased* estimate of the between-treatment contrast $\gamma_c(\nu_A)$, the expected conditional QT prolongation. Therefore, although it may seem that the two-step procedure implicitly assumes that the slopes β_A and β_P are the same, it actually handles the case of unequal slopes rather deftly in the following sense: the two-step procedure shifts, as it were, the focus of interest from QT prolongation at the reference RR interval to QT prolongation at the average RR interval (ν_A) under active treatment. One might consider the latter quantity, which can be interpreted as the "typical" QT prolongation experienced by subjects taking active treatment, to be the clinically more relevant contrast precisely when the slopes are unequal.

4.3. Two-step Procedure with Fixed Correction

In Step 1 of the two-step procedure with "fixed correction" QT measurements are corrected using one of a number of published formulae. Effectively, a fixed slope $\tilde{\beta}$ is used to correct both active and placebo data as follows:

$$\begin{aligned} y_A^{fc} &= y_A - \tilde{\beta} \cdot (x_A - x_r) \\ y_P^{fc} &= y_P - \tilde{\beta} \cdot (x_P - x_r) \end{aligned}$$

A slope of $\tilde{\beta} = 1/3$ under a log-linear model yields the correction formula of Fridericia (1920, 2003), a slope of $\tilde{\beta} = 1/2$ under a log-linear model yields the correction formula of Bazett (1920), and a slope of $\tilde{\beta} = 0.154$ under a linear model yields the Framingham correction formula (Sagie et al, 1992). Thereafter, in Step 2, the heart rate corrected data y_A^{fc} and y_P^{fc} are analyzed. Statistical inference is therefore about the parameter contrast $\gamma_f = \mathrm{E}(y_A^{fc}) - \mathrm{E}(y_P^{fc})$ which can be written as

$$\begin{aligned}
\gamma_f &= \mathrm{E}_{G_A}\left\{\mathrm{E}(y_A^{fc}|x)\right\} - \mathrm{E}_{G_P}\left\{\mathrm{E}(y_P^{fc}|x)\right\} \\
&= \mathrm{E}_{G_A}\left[\mathrm{E}\left\{y_A - \tilde{\beta}\cdot(x - x_r) \mid x\right\}\right] - \mathrm{E}_{G_P}\left[\mathrm{E}\left\{y_P - \tilde{\beta}\cdot(x - x_r) \mid x\right\}\right] \\
&= \alpha_A - \alpha_P + (\beta_A - \tilde{\beta})\cdot(\nu_A - x_r) - (\beta_P - \tilde{\beta})\cdot(\nu_P - x_r)
\end{aligned}$$

Geometrically, γ_f can be obtained as follows: project *both,* $\mathrm{E}(y_A|\nu_A)$ *and* $\mathrm{E}(y_P|\nu_P)$, onto the y-axis using the *fixed slope* $\tilde{\beta}$; γ_f is then given by the difference of the projections (Figure 1; Table 1).

It is instructive to consider γ_f for the special cases of no shift in average RR interval due to treatment ($\nu = \nu_A = \nu_P$), and of equal slopes ($\beta = \beta_A = \beta_P$), respectively: When $\nu = \nu_A = \nu_P$ we can write γ_f as

$$\begin{aligned}
\gamma_f &= \alpha_A - \alpha_P + (\beta_A - \beta_P)\cdot(\nu - x_r) \\
&= \gamma_2 = \gamma_c(\nu) = \gamma_m
\end{aligned}$$

If $\nu_A = \nu_P$, therefore, γ_f is identical to the between-treatment contrast for the two-step procedure with data driven correction, identical to the expected value of $\gamma_c(x)$, and identical to the between-treatment contrast for uncorrected QT data. These relationships suggest that for drugs that do not shift the average RR interval, heart rate correction, whether fixed or data driven, is redundant in the sense that such correction has no effect on the treatment contrast relative to the treatment contrast for uncorrected QT data.

Similarly, when $\beta = \beta_A = \beta_P$ we can write γ_f as

$$\gamma_f = \alpha_A - \alpha_P + (\beta - \tilde{\beta})\cdot(\nu_A - \nu_P) \tag{13}$$

Under the assumption of equal slopes the appropriate treatment contrast for assessment of conditional QT prolongation clearly is $\gamma_c = \alpha_A - \alpha_P$. Equation (13) then reveals how the two-step procedure with fixed correction, depending on the effect of active treatment on average RR interval, can falsely suggest conditional QT prolongation, or can mask its presence: When active treatment decreases the average RR interval ($\nu_A - \nu_P < 0$) and $\beta < \tilde{\beta}$, then $\gamma_f > \alpha_A - \alpha_P$ and conditional QT prolongation might be falsely suggested. *Vice versa*, when active treatment increases the average RR interval ($\nu_A - \nu_P > 0$) and $\beta > \tilde{\beta}$, then $\gamma_f < \alpha_A - \alpha_P$ and the presence of conditional QT prolongation might be masked.

In summary, fixed correction methods are either redundant (when the drug in question does not change average RR interval), or potentially biased (when the drug does change average RR interval) (Wang, Pan and Balch, 2008).

A case in point is the analysis published by Shah and Hajian (2003): active treatment increased the average heart rate by about 10 beats/min, and therefore decreased the average RR interval (their Table II); furthermore, the slope estimate from the one-step procedure assuming equal slopes (an assumption supported by the data) was $\hat{\beta} = 0.292$, which is only slightly smaller than Fridericia's slope of $\tilde{\beta} = 1/3$ but considerably smaller than Bazett's slope of $\tilde{\beta} = 1/2$. Data analysis using the one-step method suggested QT shortening, while the two-step method using fixed correction according to Bazett suggested QT prolongation; the two-step method using fixed correction according to Fridericia suggested QT shortening, but less extensive than the QT shortening estimated by the one-step method. Thus the results of the one-step procedure and of the two-step procedure using Bazett's correction are completely contradictory, but exactly as predicted by Equation (10) above.

4.4. Comparison of Uncorrected QT Data

For completeness we recall that in a comparison of uncorrected QT data statistical inference is about the difference of the marginal means $\gamma_m = \mu_A - \mu_P$ (1), which can expressed as in (7). For the case of no shift in average RR interval due to treatment

($\nu = \nu_A = \nu_P$), we have already pointed out in Section 4.3 that $\gamma_m = \gamma_2 = \gamma_f = \gamma_c(\nu)$. Thus, when a drug does not shift the average RR interval, marginal QT prolongation (γ_m) is the same as conditional QT prolongation at the average RR interval ($\gamma_c(\nu)$). Note, however, when $\beta_A \neq \beta_P$, an assessment of conditional QT prolongation using characteristics $\gamma_c(x_r)$ and $\gamma_{c,\max}$ can still provide information different to an assessment of marginal QT prolongation, because $\gamma_c(x_r)$ and $\gamma_{c,\max}$ characterize conditional QT prolongation at RR intervals other than ν.

In the case of equal slopes, $\beta = \beta_A = \beta_P$, γ_m is given by

$$\gamma_m = \alpha_A - \alpha_P + \beta \cdot (\nu_A - \nu_P) \tag{14}$$

Under the assumption of equal slopes the appropriate characteristic for assessment of average conditional QT prolongation clearly is $\gamma_c(x) \equiv \gamma_c = \alpha_A - \alpha_P$. Equation (14) then reveals why in general ($\nu_A \neq \nu_P$) an assessment of marginal QT prolongation – statistical inference on γ_m – can indeed provide different information from an assessment of conditional QT prolongation: β is always positive, and if active treatment decreases the average RR interval ($\nu_A < \nu_P$), then conditional QT prolongation ($\alpha_A \gg \alpha_P$) might be present in the absence of marginal QT prolongation, since $\beta \cdot (\nu_A - \nu_P)$ is negative and consequently γ_m might be small even when $\alpha_A \gg \alpha_P$. Conversely, if active treatment increases the average RR interval ($\nu_A > \nu_P$), then $\beta \cdot (\nu_A - \nu_P)$ is positive and therefore γ_m might be large (presence of marginal QT prolongation) although there is no conditional QT prolongation ($\alpha_A - \alpha_P \approx 0$).

5. PROPOSED ASSESSMENT OF QT PROLONGATION

Thorough QT trials often have a cross-over design and ECG data obtained in such trials typically have a hierarchical, cross-classified structure: each subject undergoes several treatment periods, in each treatment period multiple ECGs are obtained at various time points and within each ECG recording, QT and RR data are measured in three to four wave forms (Ring, 2009). For this reason, statistical analysis of both

marginal and conditional QT prolongation involves fitting linear or log-linear mixed models (Shah and Hajian, 2003; Dmitrienko and Smith 2003; Li et al, 2004; Ring, 2009). Furthermore, covariates such as gender might be fitted. In the following we assume that an appropriate statistical model is fitted to the data that yields statistically valid point and interval estimates of the various characteristics of marginal and conditional QT prolongation.

Schematically, assessment of conditional QT prolongation based on characteristic $\gamma_{c,\max}$ could proceed as follows:

1. Fit a linear (or log-linear) mixed model to the QT data measured under both active treatment and placebo, with, at a minimum, fixed effects as in Model (5) (" one-step procedure").

2. Report point and interval estimates for the parameters of Model (5). In addition, a point and interval estimate for the difference of slopes $(\beta_A - \beta_P)$ could be reported, as well as a test for equality of slopes.

3. Given a reference range $[x_0, x_1]$ for RR interval values, report point and interval estimates for $\gamma_c(x_0)$ and $\gamma_c(x_1)$. – Let u_0 and u_1 denote the upper limits of the two-sided 90% CI for $\gamma_c(x_0)$ and $\gamma_c(x_1)$, respectively.

   - If data are analyzed on the original scale, declare absence of conditional QT prolongation if $\max\{u_0, u_1\} \leq \delta$, where $\delta = 10$ ms is the relevant regulatory limit (ICH 2005, Section 2.2.4).

4. If data were analyzed on the logarithmic scale (log-linear mixed model was fitted), report the antilogs of the point and interval estimates for $\gamma_c(x_0)$ and $\gamma_c(x_1)$. – Let be $U_0 = \exp(u_0)$ and $U_1 = \exp(u_1)$.

   - Declare absence of conditional QT prolongation if $\max\{U_0, U_1\} \leq \Delta$, where Δ is a regulatory constant on the ratio or percent scale (to be specified) for maximum permissible QT prolongation.

Further research would be needed on the relative merits of the possible analysis scales: analysis of untransformed versus analysis of log-transformed data. If the log-linear analysis is chosen, the associated decision rule (under item 4 above) is suggested by analogy with the statistical analysis and decision rule for log-transformed pharmacokinetic data in bioequivalence studies (see, for example, Chow and Liu 2008): point estimates and associated 90% CIs for between-treatment contrasts on the logarithmic scale are back-transformed to the original scale by taking the antilog. For example, on the logarithmic scale we have a point estimate and 90% CI for the difference of conditional means $\gamma_c(x_0) = \mathrm{E}(y_A|x_0) - \mathrm{E}(y_P|x_0)$. Taking the antilog, we obtain a point estimate and 90% CI for the ratio of geometric means $R_0 = \exp\{\gamma_c(x_0)\} = \exp\{\mathrm{E}(y_A|x_0)\} / \exp\{\mathrm{E}(y_P|x_0)\}$; similarly for $\gamma_c(x_1)$. Thus analysis results for QT prolongation would be reported on the ratio or percent scale (rather than the difference scale). A value for the regulatory constant Δ (ratio scale) could be motivated as follows: since $u = 10$ ms is the maximum permissible QT prolongation on the difference scale, and assuming an average QT interval of approximately 400 ms, the maximum permissible QT prolongation on the ratio scale could be specified as $\Delta = 410/400 = 1.025$. Of course, any such limit would have to be determined by regulatory authorities based on appropriate comment and clinical input.

If, instead of $\gamma_{c,\max}$, one chooses to assess conditional QT prolongation using characteristic $\gamma_c(x_r)$, statistical inference could proceed in the same manner as above.

Statistical inference on Characteristic (10) is not possible based on a mixed model implementation of Model (5), since $\gamma_c(\nu_A)$ is a nonlinear function of parameters of both the conditional distribution of y given x, and of the marginal distribution of x. Statistical inference on $\gamma_c(\nu_A)$ could be based on the joint likelihood of y and x, written as the conditional likelihood of y given x, times the marginal likelihood of x, namely $L(y, x) = L(y|x) \cdot L(x)$. Point and interval estimates of $\gamma_c(\nu_A)$ could then be obtained using a Bayesian approach. Alternatively, a 90% CI for $\gamma_c(\nu_A)$ could be constructed using the bootstrap.

6. EXAMPLE

6.1. Data

To illustrate the proposed assessment of QT prolongation (Section 5) we consider data from a thorough QT/QTc study where healthy subjects received single doses of the following treatments: placebo, different doses of the study drug, and moxifloxacin (active control treatment). All subjects received placebo and the study drug in randomized cross-over fashion. During each of the study periods QT and RR interval data were collected at the following time points: 30 min before drug application, and at 0h30, 1h00, 1h30, 2h00, 3h00 and 4h00 after drug application. At each time point, subjects underwent 10-second ECGs, and 12 QT and RR interval measurements per subject, treatment period and time point were taken. Before data analysis, these 12 replicate QT and RR interval measurements per time point were averaged (on the logarithmic scale). For the purposes of this illustration we analysed the placebo data and data for one of the active doses (in the following labeled "Active Treatment"). In order to maintain data confidentiality, a small random noise term was added to all data points, but conclusions from the data made here remained unaffected by this measure.

6.2. Log-Linear Mixed Model

We conducted a one-step analysis of the QT interval data using a log-linear mixed model as outlined in the previous section. The basic SAS PROC MIXED code (SAS, 2004) is presented in Table 2: In the manner of Patterson, Jones and Zaffira (2005) we fitted the following fixed effects: period, treatment, timepoint, period×timepoint, treatment×timepoint, and baseline QT measurement as covariate. In addition, in order to implement the correction for RR interval, we fitted the log-transformed RR interval value as covariate; finally, to allow for different (non-parallel) slopes for the two treatments (see Equation (5)) we fitted the interaction between treatment and the RR interval covariate. The relevant SAS `MODEL` statement is presented in Table 2.

To account for within-subject correlation between QT interval data at the seven post-dose time points we fitted a repeated measures model, again in the manner of

Patterson, Jones and Zaffira (2005), but with unstructured covariance pattern because the time points in our example are not equally spaced and because the selection of potentially simpler covariance models is not the focus of this paper (`REPEATED` statement in Table 2).

Finally, in order to accommodate the individual nature of the QT/RR relationship (see, for example, Shah and Hajian, 2003) we fit random intercepts / random slopes for each subject, with an unstructured covariance pattern for this bivariate vector of random coefficients (`RANDOM` statement in Table 2).

Maximum likelihood estimates of treatment contrasts ("Active – Placebo") at each time point, and associated two-sided 90% CIs, were calculated on the logarithmic scale (`LSMEANS` statements in Table 2). The resulting point and interval estimates were back-transformed to the original scale using the antilog, to obtain point estimates and 90% CIs for the "Active / Placebo" ratios of geometric mean QT interval.

6.3. Results: Regression Slopes

Fitting the mixed model yields an average (or fixed effect) slope estimate of $\hat{\beta}_P = 0.246$ (SE = 0.025) for Placebo, and $\hat{\beta}_A = 0.181$ (SE = 0.022) for active treatment (logarithmic scale). The estimate of the difference in slopes is $\hat{\beta}_A - \hat{\beta}_P = -0.065$ (SE = 0.026), which is statistically significant (P = 0.0131) (so that this data set represents an example of a relatively large slope difference between active and placebo treatment). Thus the extent of QT prolongation depends on the RR interval: since the regression slope for active treatment is smaller than for placebo, the extent of QT prolongation increases with decreasing RR interval (that is, with increasing heart rate).

6.4. Results: Assessment of QT Prolongation

QT prolongation, expressed on the ratio scale, is summarized in Table 3 for the seven post-dose time points. Since the RR interval regression slopes are significantly different for active treatment and placebo, QT prolongation was estimated over a heart rate reference range of $[50, 70]$ beats/min (see Section 3.3 (iii)). For completeness, QT

prolongation is also presented for the conventional reference heart rate of 60 beats/min (corresponding to the conventional reference RR interval of $RR_r = 1$ s).

In this example active treatment causes significant QT prolongation, particularly within the first hour after drug administration. (The point estimates and upper limits of the 90% CIs can be compared to a value of 1.025 for maximum permissible QT prolongation on the ratio scale, as motivated in Section 5.) Furthermore, the dependence of the extent of QT prolongation on the heart rate (RR interval) is quite clear. As predicted from the relative sizes of the estimated regression slopes, QT prolongation is considerably smaller at a heart rate of 50 beats/min than at 70 beats/min. We note that both at 2h00 and 3h00 post drug administration, the results for a heart rate of 60 beats/min (conventional reference value) suggest that no significant QT prolongation occurred at those time points (upper limits of 90% CI below 1.025). However, at a heart rate of 70 beats/min, significant QT prolongation at those time points cannot be ruled out (PE of about 1.023, upper limits of 90% CI of about 1.034). At a heart rate of 50 beats/min, the 3h00 data suggest slight QT shortening.

6.5. Results: Discussion

The fact that the RR interval regression slopes are not parallel can have clinical implications. In the present example, the direction of the difference in regression slopes implies that higher heart rates are associated with increased QT prolongation. Since higher heart rates than the reference value of 60 beats/min may be typical for a target patient population, QT prolongation for a drug with RR interval regression slope significantly smaller than that of placebo must be assessed with extra care. However, it is also possible that the RR interval regression slope for active drug is significantly larger than that of placebo (we have evaluated such a data set; not shown here). In that case, QT prolongation at higher heart rates typical of a patient population is smaller than QT prolongation at the reference heart rate of 60 beats/min.

It should be noted that probably only few drugs produce slope differences as large as seen in the current example. However, the fact that such differences may exist suggests

that unequal slopes should, at least initially, be accommodated in the statistical model for analysis of QT data.

7. DISCUSSION

In this paper have suggested that an assessment of uncorrected or "absolute" QT prolongation is carried out through a between-treatment comparison of the marginal distributions of QT data; an assessment of "heart rate corrected" QT prolongation is carried out through a between-treatment comparison of the conditional distributions of QT data (conditional on RR interval). One way to carry out such a comparison of marginal and conditional distributions in practice is to compare marginal and conditional expectations, respectively. Under this framework, $\gamma_c(x) = \mathrm{E}(y_A|x) - \mathrm{E}(y_P|x)$ emerges as the appropriate treatment contrast for the assessment of conditional QT prolongation in general, and $\gamma_c(x) = \alpha_A - \alpha_P + (\beta_A - \beta_P) \cdot (x - x_r)$ under the usual assumption of a linear or log-linear model for the conditional expectations.

7.1. Choice of summary characteristic for conditional QT prolongation

Our example demonstrates that in general the assumption of equal slopes in Model (5) ($\beta_A = \beta_P$) cannot be made. In general, therefore, the extent of conditional QT prolongation, $\gamma_c(x)$, is a function of $x = f(RR)$. We have proposed three possible summary characteristics for this function: $\gamma_c(x)$ evaluated at some "reference" RR interval; the expected value of $\gamma_c(x)$; and the maximum value of $\gamma_c(x)$ over a "reference range" of RR interval values.

The expected value of $\gamma_c(x)$ (where the expectation is taken over the distribution of x under active treatment) seems an attractive choice of summary characteristic, but is associated with both statistical and conceptual difficulties. Statistically, inference on $\gamma_c(\nu_A)$ is not possible based on a mixed model implementation of Model (5), since $\gamma_c(\nu_A)$ is a nonlinear function of parameters of both the conditional distribution of y given x, and of the marginal distribution of x. Conceptually, the difficulty is that a thorough QT/QTc study might not provide a suitable estimate for $\gamma_c(\nu_A) = \mathrm{E}_{G_A}\{\gamma_c(x)\}$; trial

subjects are not a random sample of the target patient population, and therefore the distribution G_A of RR interval measurements in a trial with healthy subjects is unlikely to reflect the distribution of RR values in the target patient population. Using $\gamma_c(\nu_A)$ as characteristic for conditional QT prolongation implies that the characteristic is determined by the choice of trial subjects.

Alternatively to $\gamma_c(\nu_A)$, a regulatory authority might specify a suitable reference RR interval RR_r, so that $\gamma_c(x_r)$ is used as a summary characteristic for conditional QT prolongation. We suggest that RR_r might be chosen smaller than the conventional reference RR interval of 1 s, since a lower value might be more reflective of average or typical RR interval in the target population.

Our preferred characteristic for conditional QT prolongation is the maximum value of $\gamma_c(x)$ over a suitable "reference" range $[x_0, x_1]$ of RR interval values, where x_0 and x_1 would have to be specified by regulatory authorities. (Characteristic $\gamma_c(x_r)$ can be viewed as a special case where $x_0 = x_1 = x_r$.) For example, a reference range of $[0.857, 1.20]$ s, corresponding to heart rates from 50 to 70 beats/min, includes, at the upper end, HR values reflective of the target patient population. Because lack of QT prolongation is concluded if the upper confidence bounds of both $\gamma_c(x_0)$ and $\gamma_c(x_1)$ are below the regulatory limit, no alpha adjustment for repeated tests would be required.

7.2. One-step procedure

As we have seen, the so-called one-step procedure, by fitting Model (5) simultaneously to both active and placebo data, has at least two advantages: the procedure can provide unbiased estimates of the parameters of Model (5), and therefore the procedure can provide a "complete" characterization of conditional QT prolongation since $\gamma_c(x)$ can be estimated over any range of RR interval values of interest. Furthermore, and this point has been made by other authors, if an appropriate error model for the one-step procedure is specified, usually a linear or log-linear mixed model, then point estimates for $\gamma_c(x)$ are efficient, and the interval estimates have correct coverage.

Another advantage of the one-step procedure is that QT prolongation is assessed

directly through analysis of QT data (under a model that contains RR interval as a covariate): it is not necessary to calculate QTc, the heart rate corrected QT data. The choice of various QT correction methods, e.g. data driven with "population" slopes, data driven with "individual" slopes, or data driven with shrinkage of individual slopes, is replaced by the equivalent choice of an appropriate mixed model for the QT data, fixed and random effects. However, unlike the various two-step procedures, the one-step procedure provides a standard statistical framework for proper model selection.

7.3. Two-step procedure with data driven correction

We have shown that the two-step procedure with data driven correction provides an unbiased estimate for $\gamma_c(\nu_A)$, that is, of the expected conditional QT prolongation, where the expectation is taken with respect to the distribution of $x = f(RR)$ under active treatment. We have argued that this contrast in principle is meaningful, and might indeed be of greater clinical interest than $\gamma_c(x_r)$, particularly when the reference RR interval is poorly chosen. However, as pointed out above, a thorough QT/QTc study in healthy volunteers probably is not the ideal setting for estimating $\gamma_c(\nu_A)$. Furthermore, interval estimates of the treatment contrast from the two-step procedure do not have the correct coverage (Dmitrienko and Smith, 2003): whether population or individual data driven correction methods are used, multiple QT values are corrected using the same slope estimate. The resulting correlation between QTc data (and reduction of error degrees of freedom) is not taken into account when the data are analyzed. Consequently, confidence intervals for $\gamma_c(\nu_A)$ from the two-step procedure have coverage smaller than their nominal coverage so that the two-step procedure is not valid statistically. Two methods for proper statistical inference about $\gamma_c(\nu_A)$ were sketched in Section 5.

7.4. Two-step procedure with fixed correction

The treatment contrasts for the various fixed correction procedures, namely $\gamma_f = \mathrm{E}(y_A^{fc}) - \mathrm{E}(y_P^{fc})$, are not appropriate characteristics for conditional QT prolongation.

Depending on the effect of active treatment on average RR interval, the two-step procedure with fixed correction can falsely suggest conditional QT prolongation, or mask its presence. QT "correction" according to Bazett is particularly notorious in this regard because Bazett's regression slope tends to be considerably larger than regression slopes estimated from modern data bases (see, for example, Dmitrienko and Smith, 2003; Shah and Hajian, 2003). Bazett's procedure will falsely suggest QT prolongation for drugs that increase heart rate, and will mask QT prolongation for drugs that decrease heart rate. (No harm is done only when active treatment does not shift average heart rate, but only because in this case correction has no effect and is redundant.) The consequences of falsely declaring an unsafe drug safe, or of falsely declaring a safe drug unsafe, are obviously serious. All fixed correction methods are potentially biased, and the bias of Bazett's method is the most severe. Fixed correction may be appropriate when single or sparse ECG data are evaluated, such as in routine clinical practice; however, we think that in the statistical analysis of thorough QT studies fixed correction methods should be abandoned.

7.5 Conclusion

Conditional QT prolongation $\gamma_c(x)$ in general is a function of RR interval or heart rate. The various procedures for assessment of QT prolongation have different requirements, for example, regarding the number of sampling points and number of subjects. From the point of view of an adequate parametric characterization of QT prolongation, however, we suggest fitting a linear or log-linear mixed model in an implementation of the one-step procedure, and allowing, in general, for non-parallel regression slopes for the different treatments.

ACKNOWLEDGMENT

We thank two referees for valuable comments.

REFERENCES

Bazett, J.C. (1920). An analysis of time relations of electrocardiograms. *Heart* 7: 353-367.

Chow, S. C., Liu, J. P. (2008). *Design and Analysis of Bioavailability and Bioequivalence Studies. Third Edition.* New York: CRC Press.

Dmitrienko, A., Smith, B. (2003). Repeated-measures models in the analysis of QT interval. *Pharmaceutical Statistics* 2: 175–190.

Fridericia, L.S (1920). Die Systolendauer im Elektrokardiogramm bei normalen Menschen und bei Herzkranken. Teil I: Beziehung zwischen der Pulsfrequenz und der Dauer des Ventrikelelektrokardiogramms bei normalen Menschen in der Ruhe. *Acta Med Scand* 53: 469–486.

Fridericia, L.S. (2003). The duration of systole in an electrocardiogram in normal humans and in patients with heart disease (English translation of Fridericia, 1920). *Annals of Noninvasive Electrocardiology* 8: 343–351.

ICH (2005). *The Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs.* ICH Harmonised Tripartite Guideline E14, Step 4. Available at http://www.ich.org/.

Li L., Desai, M., Desta, Z., Flockhart, D. (2004). QT analysis: a complex answer to a "simple" problem. *Statistics in Medicine* 23: 2625–2643

Malik, M., Fářbom, P., Batchvarov, V., Hnatkova, K., Camm, A.J. (2002). Relation between QT and RR intervals is highly individual among healthy subjects: implications for heart rate correction of the QT interval. *Heart* 87: 220–228.

Patterson, S.D., Jones, B., Zariffa, N. (2005). Modeling and interpreting QTc prolongation in clinical pharmacology studies. *Drug Information Journal* 39: 437–445.

Ring, A. (2009). Statistical models for the heart rate correction of the QT Interval. *Statistics in Medicine*, in press.

Sagie, A., Larson, M.G., Goldberg, RJ., Bengtson, J.R., Levy, D. (1992) An improved method for adjusting the QT interval for heart rate (the Framingham

Heart Study). *American Journal of Cardiology* 70: 797–801.

SAS Institute Inc. (2004). *SAS/STAT 9.1 User's Guide. Volume 4.* Cary, NC: SAS Institute Inc.

Schall, R., Luus, H.G. (1993). On population and individual bioequivalence. *Statistics in Medicine* 12: 1109–1124.

Shah, A., Hajian, G. (2003). A maximum likelihood approach for estimating the QT correction factor using mixed effects model. *Statistics in Medicine* 22: 1901–1909.

Wang, Y., Pan, G., Balch, A. (2008). Bias and variance evaluation of QT interval correction methods. *Journal of Biopharmaceutical Statistics* 18: 427-450.

Table 1. Geometric Derivation of Characteristics for QT Prolongation

	Method	Characteristic	Projection Slope for	
			$\mathrm{E}(y_A\|\nu_A)$	$\mathrm{E}(y_P\|\nu_P)$
1	Marginal QT Prolongation	γ_m	zero	zero
2	One Step Method	$\gamma_c(x_r)$	β_A	β_P
3	Two Step Method – data driven	$\gamma_c(\nu_A)$	β_P	β_P
4	Two Step Method – fixed	γ_f	$\tilde{\beta}$	$\tilde{\beta}$

Table 2. SAS Code for One-Step Analysis of QT Data (Example of Section 6)

```
PROC MIXED;
   CLASS subject treat time period;
   MODEL y = baseline period treat time period*time treat*time
             x treat*x / DDFM=KR;
   RANDOM int x / SUBJECT=subject TYPE=UN;
   REPEATED time / SUBJECT=subject*treat TYPE=UN;
   ESTIMATE 'slope A' x 1 treat*x 1 0 / CL;
   ESTIMATE 'slope P' x 1 treat*x 0 1 / CL;
   ESTIMATE 'slopediff A - P' treat*x 1 -1 / CL;
   LSMEANS treat*time / DIFF AT x=7.09008 CL ALPHA=0.1;
   LSMEANS treat*time / DIFF AT x=6.90776 CL ALPHA=0.1;
   LSMEANS treat*time / DIFF AT x=6.75360 CL ALPHA=0.1;
RUN;
```

subject: Subject; treat: Treatment; time: Time Point; period: Study Period

y: QT interval measurement [ms] after logarithmic transformation: y=ln(QT)

x: RR interval measurement [ms] after logarithmic transformation: x=ln(RR)

"DIFF AT x=7.09008": LSMEANS differences calculated at x=ln(1000*60/50), that is at heart rate of 50 beats/min or RR interval of 1000*60/50=1200 ms; similarly for "DIFF AT x=6.90776" (RR interval of 1000*60/60=1000 ms) and "DIFF AT x=6.75360" (RR interval of 1000*60/70=857 ms)

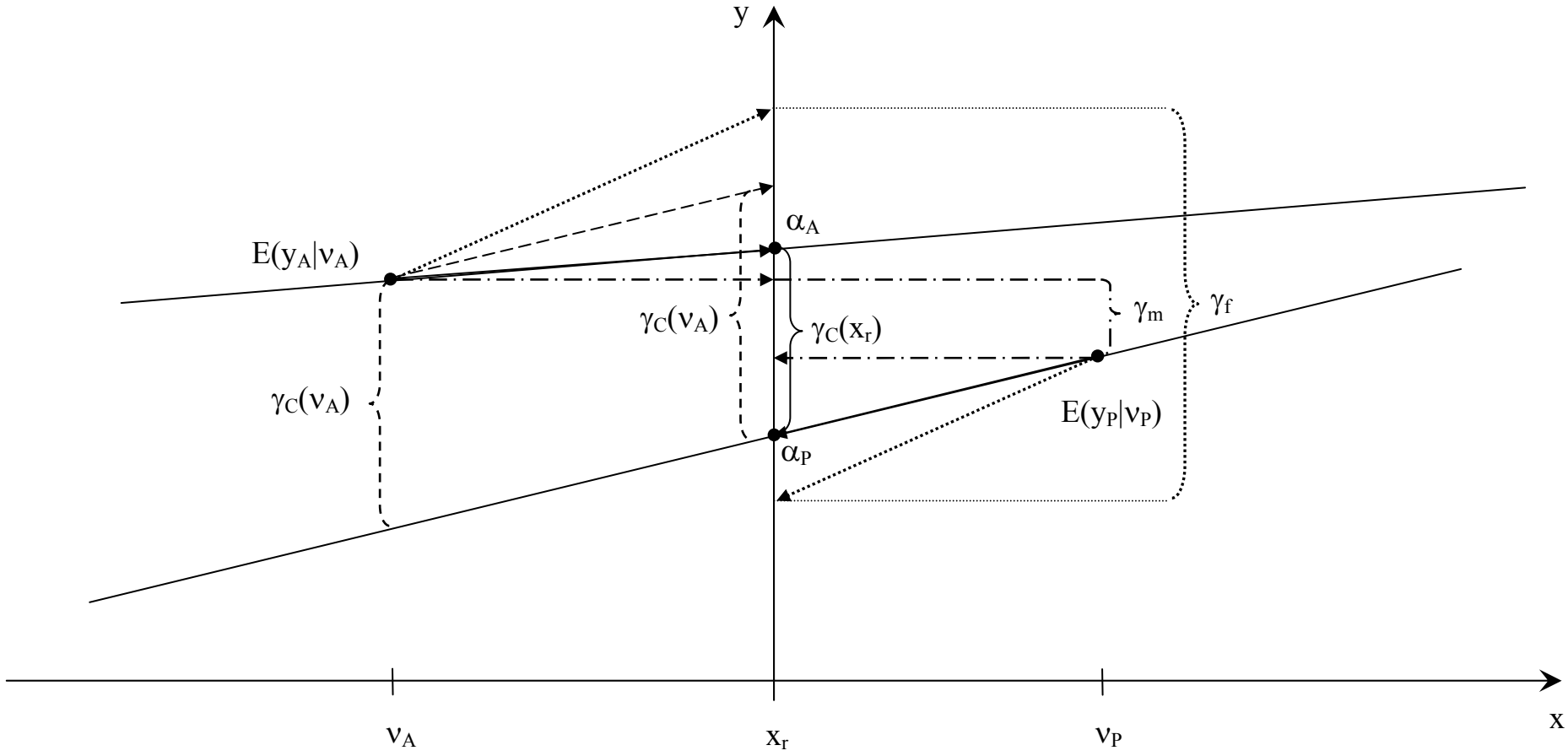
A: Active Treatment

P: Placebo Treatment

Table 3. Assessment of QT Prolongation (Example of Section 6)

Heart Rate	Time	Geom. Mean QT Int. [ms]		Ratio "Active/Placebo"	
[beats/min]		Active	Placebo	PE	90% CI
50	0h30	419.8	411.2	1.0209	1.0094 – 1.0325
	1h00	422.8	414.7	1.0198	1.0088 – 1.0309
	1h30	418.6	415.5	1.0072	0.9955 – 1.0191
	2h00	415.2	414.7	1.0013	0.9896 – 1.0131
	3h00	413.5	413.6	0.9997	0.9872 – 1.0124
	4h00	410.7	414.7	0.9903	0.9779 – 1.0029
60	0h30	406.2	393.2	1.0330	1.0242 – 1.0419
	1h00	409.1	396.4	1.0319	1.0239 – 1.0399
	1h30	405.0	397.3	1.0192	1.0103 – 1.0282
	2h00	401.7	396.5	1.0132	1.0044 – 1.0220
	3h00	400.1	395.5	1.0116	1.0015 – 1.0218
	4h00	397.3	396.5	1.0021	0.9924 – 1.0118
70	0h30	395.0	378.5	1.0434	1.0319 – 1.0549
	1h00	397.8	381.7	1.0423	1.0316 – 1.0530
	1h30	393.8	382.6	1.0294	1.0182 – 1.0408
	2h00	390.7	381.8	1.0233	1.0124 – 1.0344
	3h00	389.1	380.8	1.0218	1.0096 – 1.0340
	4h00	386.4	381.8	1.0121	1.0006 – 1.0238

PE: Point Estimate
CI: Confidence Interval


y
α_A
$E(y_A|\nu_A)$
$\gamma_C(\nu_A)$
$\gamma_C(x_r)$
γ_m
γ_f
$\gamma_C(\nu_A)$
$E(y_P|\nu_P)$
α_P
x
ν_A
x_r
ν_P

SOME ASPECTS OF COMMON SINGULAR SPECTRUM ANALYSIS AND COINTEGRATION OF TIME SERIES.

D.G. Nel and H. Viljoen, Department of Statistics and Actuarial Science,

Stellenbosch University, South Africa

Abstract: Singular spectrum analysis (SSA) is a time series modelling technique where an observed time series is unfolded into the column vectors of a Hankel structured matrix, known as a trajectory matrix. For deterministic series the column vectors of the trajectory matrix lie on a single R-flat. Singular value decomposition (SVD) can be used to find the orthonormal base vectors of the linear subspace parallel to this R-flat. SSA is useful to model time series with complex cyclical patterns that increase over time.

Common singular spectrum analysis was investigated by Viljoen and Nel (2009). In this paper the method is briefly discussed and the similarities with cointegration is investigated, the most important that time series when sharing similarities identified by co-integration, shares a common R-flat of dimension r, which need to be determined. Common singular spectrum analysis provides the methodology to identify r by using the common principal component (CPC) approach of Flury (1988). CSSA decomposes the different original time series into the sum of a common small number of components which are related to common trend and oscillatory components and noise. The similarities and differences between cointegration and CSSA are studied simulating several different scenarios.

Keywords: Co-integration, hierarchical approach, singular spectrum analysis, Singular value decomposition, Common principal components

1. Introduction

Singular spectrum analysis of time series was introduced by Broomhead and King (1986a, 1986b). Extensive literature exists regarding the development of the method since then, for example E.G Buchstaber (1994), J.B. Elsner and A.A. Tsonis (1996), Danilov (1997) and Golyandina et al (2001). These methods are also known as caterpillar methods (Caterpillar 3.30) and are used to study the structure of time series. The purpose is to unfold a time series into a trajectory matrix whose singular values are then determined to reconstruct a smoother time series which can be used for explaining structure and for forecasting.

Flury (1984) derived methods to determine common principal components of several symmetric matrices, usually covariance matrices in multivariate analysis, under the assumption that these matrices have a common principal component (CPC) structure.

CSSA is a method utilizing the methodology of Flury (1988) to determine the common base vectors which spans the *r*-dimensional manifold (or R-flat) suspected to be common to both series. Co-integration analysis is the most frequently used method to study such common structures among time series (Engle and Granger (1987), Johansen (1988), (1991) and Wei (2006)). The purpose of this paper is to discuss relationships between CSSA for only two time series and co-integration.

Singular spectrum analysis of time series (SSA) is discussed briefly in section 2 and common principal component analysis (CPC) and partial common principal component analysis (CPC(r)) in section 3. In section 4 the common singular spectrum analysis method is introduced. In section 5 a heuristic method and a hierarchical method are discussed to determine the dimensionality of the common supporting linear subspace or R-flat. An example is presented in section 6 where two series share

common features. In section 7 several simulated studies are done to compare CSSA to co-integration.

2. Singular spectrum analysis (SSA)

In SSA methodology an observed time series X_t is unfolded into the column vectors of a matrix $\boldsymbol{X}:(\tau \times n); n = T - \tau + 1$:

$$\boldsymbol{X} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ x_2 & x_3 & \dots & x_{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_\tau & x_{\tau+1} & \dots & x_T \end{bmatrix}.$$

The Hankel matrix $\boldsymbol{X}$ is called the trajectory matrix and the number τ of rows of $\boldsymbol{X}$ is referred to as the window length. The process to stack the elements of a time series into a Hankel matrix is called the hankelization of a time series. The window length is the dimension of the Euclidean space into which the time series is unfolded and the choice of τ is restricted to $2 \le \tau \le [(T+1)/2]$, where the notation [b] indicates the integer part of a fraction b. Golyandina et al (2001), section 1.6, give general outlines for the choice of the window length. The window length should be chosen as a multiple of the periodicity but not exceeding half of the time series.

Buchstaber (1994) noted that if a deterministic time series is unfolded into the column vectors of a trajectory matrix, all the column vectors of the trajectory matrix lie on a single R-flat (H_r). The term R-flat indicates a r-dimensional manifold which does not necessarily passes through the origin.

Venter (1998) illustrated how the R-flat H_r can be considered as a combination of a shift vector $\underline{b}$ and parallel linear subspace $\Im_r$, where $\Im_r = span(\underline{v}_1, \underline{v}_2, \dots, \underline{v}_r)$. The R-flat H_r is then defined as:

$$H_r = \left\{ \underline{x} : \underline{x} = \underline{b} + a_1 \underline{v}_1 + \ldots a_r \underline{v}_r ; a_i \in \mathbb{R}; \underline{v} \in \mathbb{R}^\tau ; \tau \geq r+1 \right\},$$

where the vectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_r$ form a base for the parallel linear subspace and are therefore linearly independent.

For an observed time series the singular value decomposition (SVD) is used to find a reconstructed signal series $\tilde{f}_t$, which can be used for forecasting. (Golyandina et al. 2001). First it is necessary to find a low-rank approximation of the trajectory matrix $\boldsymbol{X}$, say matrix $\boldsymbol{Y}$ from which the series $\tilde{f}_t$ can be obtained.

The singular value decomposition (SVD) of a $(\tau \times n)$ matrix $\boldsymbol{X}$, of rank r, is defined by $\boldsymbol{X} = \boldsymbol{B}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}'$, where $\underset{(\tau \times r)}{\boldsymbol{B}} = \left[\underline{b}_1 \quad \underline{b}_2 \quad \ldots \quad \underline{b}_r \right]$ is the matrix of normalized eigenvectors of the column space of $\boldsymbol{X}$, $\underset{(n \times r)}{\boldsymbol{U}} = \left[\underline{u}_1 \quad \underline{u}_2 \quad \ldots \quad \underline{u}_r \right]$ is the matrix of normalized eigenvectors of the row space of $\boldsymbol{X}$ and matrix $\boldsymbol{\Lambda}^{\frac{1}{2}} : (r \times r)$ is a diagonal matrix of singular values of the matrix $\boldsymbol{X}$. The matrix $\boldsymbol{X}\boldsymbol{X}'$ is then diagonalized as $\boldsymbol{B}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{B} = \boldsymbol{\Lambda}$.

In SSA, SVD is used to diagonalize the symmetric $\tau \times \tau$ scatter matrix $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}'$ where $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \underline{\bar{x}}\,\underline{j}'$ the centred trajectory matrix with $\underline{\bar{x}}$ the vector of means and $\underline{j}'$ the transpose of the unit vector. The vector of means $\underline{\bar{x}}$ is used as the shift vector of the r-flat H_r. The result is given by $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}' = \boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}'$. The SVD of the matrix $\tilde{\boldsymbol{X}}$ will tend to possess r larger singular values and $(\tau - r)$ smaller singular values which can be ascribed to noise in the time series. Suppose that the trajectory matrix is of rank $d \leq \min(\tau, n)$ then the problem statement in SSA is to find a low-rank approximating matrix $\boldsymbol{Y}$ of rank $r \leq d$ for the trajectory matrix $\boldsymbol{X}$. The first r eigenvectors need to be

selected from the columns of matrix $\boldsymbol{B}$. The choice of r can be assisted by using a scree plot or using phase state portraits. (Vautard et al.1992).

Once the choice of r is made, the projection matrix of the parallel linear subspace spanned by the r leading eigenvectors $\boldsymbol{P}_{H_r} = \boldsymbol{B}\boldsymbol{B}'$ is formed. Then the closest r-flat (H_r) to the column vectors of the trajectory matrix is given by $\boldsymbol{Y} = \underline{\bar{x}}\,\underline{j}' + \boldsymbol{P}_{H_r}\tilde{\boldsymbol{X}}$. The de-hankelization operation (Buchstaber 1994) performed on $\boldsymbol{Y}$ will yield the reconstructed signal series $\tilde{f}_t$ as follows.

$$\tilde{f}_t = \begin{cases} \dfrac{1}{s}\displaystyle\sum_{i=1}^{s} Y_{i,s-i+1} & \text{for } 1 \le s \le \tau \\ \dfrac{1}{\tau}\displaystyle\sum_{i=1}^{\tau} Y_{i,s-i+1} & \text{for } \tau \le s \le n \\ \dfrac{1}{N-s+1}\displaystyle\sum_{i=1}^{N-s+1} Y_{i+s-n,n-i+1} & \text{for } n \le s \le N \end{cases}$$

The reconstructed smoother series can then be used for forecasting.

3. Common principal component and partial common principal component models

Flury (1984) introduced common principal components of two or more symmetric covariance matrices as a possible model in a hierarchy of models to explain heteroscedasticity among covariance matrices. In this paper the concepts of common principal components CPC and partial common principal components CPC(r) will be used and are briefly described for clarity.

Two (or more) symmetric matrices $\boldsymbol{S}_1 : p \times p$ and $\boldsymbol{S}_2 : p \times p$ have common principal components if an orthogonal matrix $\boldsymbol{\beta} = \left[\underline{b}_1 \quad \underline{b}_2 \quad \ldots \quad \underline{b}_p\right]$ exists such that $\boldsymbol{\beta}'\boldsymbol{S}_i\boldsymbol{\beta} = diag\left(l_1^{(i)},\ldots,l_r^{(i)},l_{r+1}^{(i)},\ldots,l_p^{(i)}\right)$, $i = 1,2$ meaning that:

$$l_j^{(i)} = \underline{b}_j'\boldsymbol{S}_i\underline{b}_j \text{ for } j = 1,\ldots,p\,.$$

If CPC structure exists, the FG-algorithm of Flury and Gautschi (1986) can be used to find the maximum likelihood estimates of the eigenvalues and the common eigenvectors in the orthogonal matrix $\boldsymbol{\beta}$.

If two matrices $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ exist which contains r common eigenvectors and diagonalize the scatter matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, e.g. they are of the form

$$\beta_i = \left[\boldsymbol{B} \quad \boldsymbol{B}_i^{(2)} \right] = \left[\underline{b}_1 \quad \dots \quad \underline{b}_r \quad \underline{b}_{r+1}^{(i)} \quad \dots \quad \underline{b}_\tau^{(i)} \right], \; i = 1,2 \;\; ,$$

then $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ satisfy a partial common principal components or CPC(r) model. Thus r of the eigenvectors forming the submatrix $\boldsymbol{B}$ of dimension $\tau \times r$ are common while the submatrices $\boldsymbol{B}_i^{(2)}$: $\tau \times (\tau - r)$ are specific. Flury (1988) Chapter 6 described the estimation of these matrices using the Flury-Gautschi algorithm. The CPC(r) model or initially the CPC model (actually a CPC(τ-1) model) can be used to determine the common eigenvectors $\boldsymbol{B} : \tau \times r$ to be used in extending the SSA model to CSSA. The estimation of the matrices is rather complicated, but an approximate estimation procedure of these matrices $\boldsymbol{B}$ and $\boldsymbol{B}_i^{(2)}$, $i = 1,2$ is described in Flury (1988) p.130. This approach is recommended rather than solving the complete system of equations as described by Flury (1988) p.203, section 9.7.

4. The Common Singular Spectrum Analysis (CSSA) method

Two time series X_{t_1} and X_{t_1} of equal length and observed over the same period of time are considered. It is assumed that common properties exist which explain the variation in both time series. It is assumed that both series can be described by models of the form:

$$X_{t_i} = f_{t_i} + \varepsilon_{t_i}, \;\; t_i = 1, \dots, T, \; i = 1,2$$

where X_{t_i} denotes the observed time series of length T, f_{t_i} the deterministic time series or signal series of length T and ε_{t_i} the white noise series.

In terms of SSA methodology this means that a common R-flat H_r defined by:

$$H_r = \left\{ \underline{x} : \underline{x} = \underline{b} + a_1 \underline{v}_1 + \ldots a_r \underline{v}_r ; a_i \in \mathbb{R}; \underline{v} \in \mathbb{R}^\tau ; \tau \geq r+1 \right\},$$

may exist for both series. Both the dimensionality r and the common base vectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_r$ spanning the common parallel linear subspace need to be determined. The CPC and CPC(r) methods of Flury (1988) provide the necessary methodology to accomplish this. The CSSA method is described in Viljoen and Nel (2009) and it is briefly given here again for illustrative purposes: Both series are unfolded into two trajectory matrices and the centered trajectory matrices $\tilde{\boldsymbol{X}}_1 : \tau \times n$ and $\tilde{\boldsymbol{X}}_2 : \tau \times n$ are formed similar to the SSA method. The symmetric matrices $\boldsymbol{S}_1 = \tilde{\boldsymbol{X}}_1 \tilde{\boldsymbol{X}}_1^{'} : \tau \times \tau$ and $\boldsymbol{S}_2 = \tilde{\boldsymbol{X}}_2 \tilde{\boldsymbol{X}}_2^{'} : \tau \times \tau$ are formed and the FG algorithm (Flury and Gautschi (1986)) applied on the matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ to determine a matrix $\boldsymbol{B} : \tau \times r$ using the CPC model, which yields a matrix

$\boldsymbol{\beta} = \left[\underline{b}_1 \quad \ldots \quad \underline{b}_r \quad \underline{b}_{r+1} \quad \ldots \quad \underline{b}_\tau \right] = \left[\boldsymbol{B} \quad \boldsymbol{B}^{(2)} \right]$ where $\boldsymbol{B} : \tau \times r$ and $\boldsymbol{B}^{(2)} : \tau \times (\tau - r)$,

such that

$$\boldsymbol{\beta}' \boldsymbol{S}_i \boldsymbol{\beta} = diag \left[l_1^{(i)} \quad \cdots \quad l_r^{(i)} \quad l_{r+1}^{(i)} \quad \cdots \quad l_\tau^{(i)} \right], \; i = 1, 2 .$$

The first r common eigenvectors $\boldsymbol{B} = \left[\underline{b}_1, \cdots, \underline{b}_r \right] : \tau \times r$ are of particular interest in CSSA and the remaining eigenvectors $\underline{b}_{r+1}, \cdots, \underline{b}_\tau$ are assumed to be associated with noise in both series. The matrix $\boldsymbol{B} : \tau \times r$ is the matrix of common eigenvectors spanning the column space of both matrices $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$ such that

$\tilde{\boldsymbol{X}}_1 = \mathbf{B}' \boldsymbol{\Lambda}_1^{\frac{1}{2}} \boldsymbol{U}_{X_1}$ and $\tilde{\boldsymbol{X}}_2 = \mathbf{B}' \boldsymbol{\Lambda}_2^{\frac{1}{2}} \boldsymbol{U}_{X_2}$. Note that the row space of both matrices is still spanned by different U-matrices.

The remainder of the CSSA algorithm is similar to SSA: Form the projection matrix $\boldsymbol{P}_{H_r} = \boldsymbol{B}\boldsymbol{B}'$, where $\boldsymbol{B} : \tau \times r$ and form the projected matrices $\boldsymbol{Y}_i = \underline{x}_i \, \underline{j}_n' + \boldsymbol{P}_{H_r} \tilde{\boldsymbol{X}}_i$, $i = 1,2$. Using the dehankelization operation on the projected matrices $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ will yield the reconstructed series $\tilde{f}_{t_1}$ and $\tilde{f}_{t_2}$.

The matrix $\boldsymbol{B} : \tau \times r$ can be determined more accurately by fitting a CPC model firstly to the scatter matrices as described above and then fitting a CPC(r) model, with $\boldsymbol{\beta}_i = \left[\boldsymbol{B} \quad \boldsymbol{B}_i^{(2)} \right] = \left[\underline{b}_1 \quad \ldots \quad \underline{b}_r \quad \underline{b}_{r+1}^{(i)} \quad \ldots \quad \underline{b}_\tau^{(i)} \right]$, *i=1,2* which essentially specify that the matrices $\boldsymbol{B}_i^{(2)} : \tau \times (\tau - r)$ can be specific to the two time series, while the base vectors of the R-flat in $\boldsymbol{B} : \tau \times r$ are common. If the CPC(r) method was used to determine $\boldsymbol{B} : \tau \times r$ the CSSA method will be refered to as the CSSA(r) method. Fitting the CPC model serves as a first step in the CSSA(r) method.

5. Choosing the dimension *r*

The number r of eigenvectors in $\boldsymbol{B} : \tau \times r$ needed from the columns of the matrix $\boldsymbol{\beta} : \tau \times \tau$ for the CSSA method can be selected using either a heuristic approach from CSSA or a hierarchical selection procedure from CSSA(r).

The heuristic approach is as follows: Arrange the eigenvectors in $\boldsymbol{\beta} : \tau \times \tau$ to yield the eigenvalues of $\boldsymbol{S}_1$ in descending order: e.g. $l_1^{(1)} \geq l_2^{(1)} \geq \ldots \geq l_d^{(1)} \geq \ldots \geq l_\tau^{(1)}$. Now apply the same sequencing to the eigenvalues of the matrix $\boldsymbol{S}_2$ to determine the index d such that $l_1^{(2)} \geq l_2^{(2)} \geq \ldots \geq l_d^{(2)}$ but the remaining eigenvalues $l_{r+1}^{(2)}, l_{r+2}^{(2)}, \ldots, l_\tau^{(2)}$ of $\boldsymbol{S}_2$ need not be ordered. The dimension r can then at most be *d.* Alternatively the eigenvalues of matrix $\boldsymbol{S}_2$ can be sorted in descending order and then apply this sequencing to the

eigenvalues of matrix $\boldsymbol{S}_1$. If the values for *d* differ from these two methods then *d* will be the minimum value. The square roots of the eigenvalues $l_1^{(1)} \geq l_2^{(1)} \geq \ldots \geq l_d^{(1)}$ and $l_1^{(2)} \geq l_2^{(2)} \geq \ldots \geq l_d^{(2)}$ are now the singular values of the common singular value decomposition of $\tilde{\boldsymbol{X}}_1 : \tau \times n$ and $\tilde{\boldsymbol{X}}_2 : \tau \times n$.

If *d* is very small, e.g. $d = 1$ then $r \leq 1$ too and the scree plots of the second series according to its own ordering and according to the ordering of the first series will be quite different. Thus a small value of *r* or immediate differences between the plots in this scree plot**,** is a clear indication that the CSSA structure is not applicable for the two time series and that no common R-flat exist. Due to the magnitude of the eigenvalues, the log-eigenvalues are sometimes rather plotted which gives a better illustration.

The hierarchical selection procedure is described in Viljoen and Nel (2009), where the CSSA(r) approach is used for this purpose. If *r* is assumed known, then a submatrix of common eigenvectors $\boldsymbol{B} : \tau \times r$ exist in both matrices $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$. The value of *r* is now selected using a similar hierarchical selection procedure as described by Flury (1988) p.148 – 151.

Two time series models CSSA$(r+1)$ and CSSA(r) for *r=1,2, ...d* are compared by fitting CPC$(r+1)$ and CPC(r) models to the scatter matrices and comparing the improvement in the χ^2- statistic, $\chi^2_{r+1/r} = \sum_{i=1}^{2} n_i \left(\log \left| \boldsymbol{S}_i^{(r+1)} \right| - \log \left| \boldsymbol{S}_i^{(r)} \right| \right)$ (Flury 1988, p150 eqn(1.1)) which is asymptotically distributed χ^2 with degrees of freedom, $df = \tau - (r+1)$ where $\boldsymbol{\Lambda}_i^{(r)} = diag\left(\boldsymbol{\beta}_i' \boldsymbol{S}_i \boldsymbol{\beta}_i \right)$, $\boldsymbol{S}_i^{(r)} = \boldsymbol{\beta}_i \boldsymbol{\Lambda}_i^{(r)} \boldsymbol{\beta}_i'$, and $\boldsymbol{\beta}_i = \begin{bmatrix} \underset{(r)}{\boldsymbol{B}} & \underset{(\tau - r)}{\boldsymbol{B}_i^{(2)}} \end{bmatrix}$, . *i=1,2.* For $r = 0$, CPC(1) is compared to unrelated $\boldsymbol{S_i}$ with

$\chi^2_{\mathrm{CPC}(1)} = \chi^2_1 = \sum_{i=1}^{2} n_i \left(\log \left| \boldsymbol{S}_i^{(1)} \right| - \log \left| \boldsymbol{S}_i \right| \right)$ and associated degrees of freedom $df = \tau - 1$, where $\boldsymbol{S}_i$ denotes the original scatter matrices. The following table is useful to measure the change. (Viljoen and Nel (2009)).

$r+1$	**Time series**	**Scatter matrices**	(*df*)	$\chi^2_{r+1/r}$
1	$\mathrm{CSSA}(1)$ vs unrelated	$\mathrm{CPC}(1)$ vs unrelated	$\tau - 1$	$\chi^2_{CPC(1)}$
2	$\mathrm{CSSA}(2)$ vs $\mathrm{CSSA}(1)$	$\mathrm{CPC}(2)$ vs $\mathrm{CPC}(1)$	$\tau - 2$	$\chi^2_{2/1}$
3	$\mathrm{CSSA}(3)$ vs $\mathrm{CSSA}(2)$	$\mathrm{CPC}(3)$ vs $\mathrm{CPC}(2)$	$\tau - 3$	$\chi^2_{3/2}$
-	--------	----	---	---
$\tau - 1$	$\mathrm{CSSA}(\tau - 1)$ vs $\mathrm{CSSA}(\tau - 2)$	$\mathrm{CPC}(\tau - 1)$ vs $\mathrm{CPC}(\tau - 2)$	1	$\chi^2_{\tau-1/\tau-2}$

Table 1 Comparing CSSA models

The value of *r* is detected where the ratio $\chi^2_{r+1/r} / df$ is closest to 1 (one) among the possible values $1 \le r \le \tau - 2$, since the CPC(τ-1) model coincides with the CPC model. The values of *r* where $1 \le r \le d$ may be more realistic to choose from since the dimension *r* of the R-flat should be much less than τ.

It was noted that if two time series do not share a common R-flat, the sequencing of the eigenvalues or log eigenvalues are immediately quite different and the value of *d* very small. The following scree plot of the log eigenvalues of series 2 sorted according to the sequence of the series 1 eigenvalues and not, illustrates this situation.

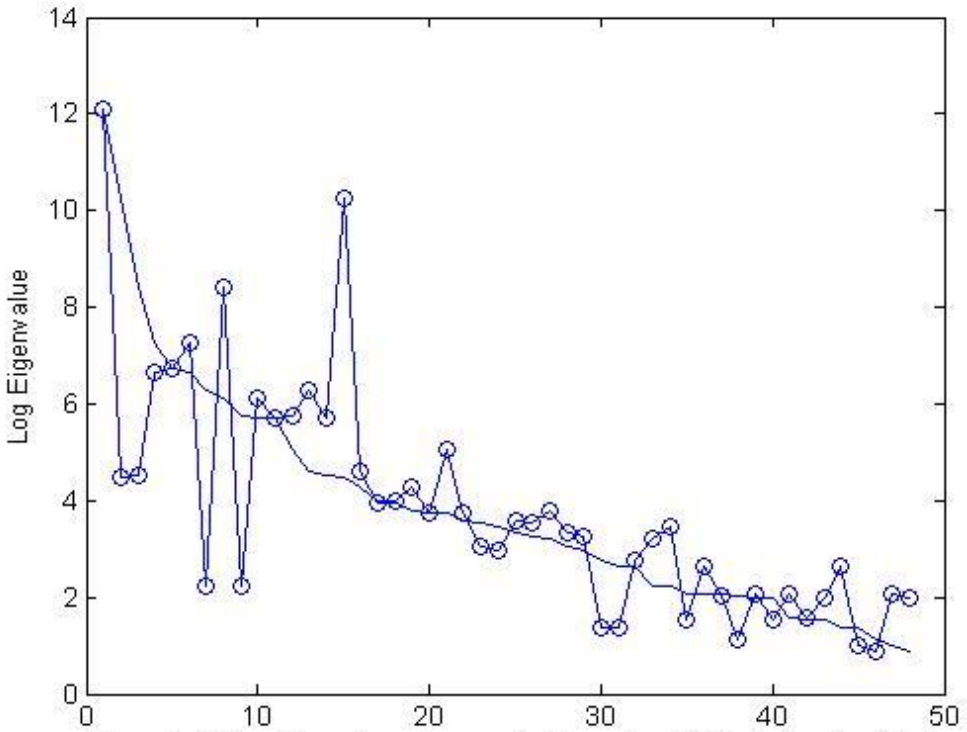


Figure 1: Series 2 log eigenvalues sorted to series 1 (dots) otherwise (line)

When no common R-flat exists, the hierarchical approach immediately gives high chi-square values and ratios $\chi^2_{r+1/r}/df$, close to 1, which the following table illustrates.

The hierarchical approach:

$r+1$	**Time series**	$\chi^2_{r+1/r}$	(df)	$\chi^2_{r+1/r}/df$
1	$\mathrm{CSSA}(1)$ vs unrelated	44.45	47	0.95
2	$\mathrm{CSSA}(2)$ vs $\mathrm{CSSA}(1)$	45.11	46	0.98
3	$\mathrm{CSSA}(3)$ vs $\mathrm{CSSA}(2)$	52.18	45	1.16
4	$\mathrm{CSSA}(4)$ vs $\mathrm{CSSA}(3)$	21.67	44	0.49
5	$\mathrm{CSSA}(5)$ vs $\mathrm{CSSA}(4)$	11.01	43	0.26
6	$\mathrm{CSSA}(6)$ vs $\mathrm{CSSA}(5)$	7.80	42	0.19
7	$\mathrm{CSSA}(7)$ vs $\mathrm{CSSA}(6)$	43.74	41	1.07
8	$\mathrm{CSSA}(8)$ vs $\mathrm{CSSA}(7)$	16.41	40	0.41
9	$\mathrm{CSSA}(9)$ vs $\mathrm{CSSA}(8)$	56.32	39	1.44
10	$\mathrm{CSSA}(10)$ vs $\mathrm{CSSA}(9)$	19.02	38	0.50
--	--	--	--	--

Table 2 Comparing CSSA models in a no common R-flat case.

In section 6 an example is given where two times series will be analyzed by using CSSA sharing common properties and a common R-flat of dimension r.

6. Examples

Example 1

The well known Lydia Pinkham annual advertising and sales dataset from 1907 to 1960 is used as a first example. (Data W12 of Wei 2006).

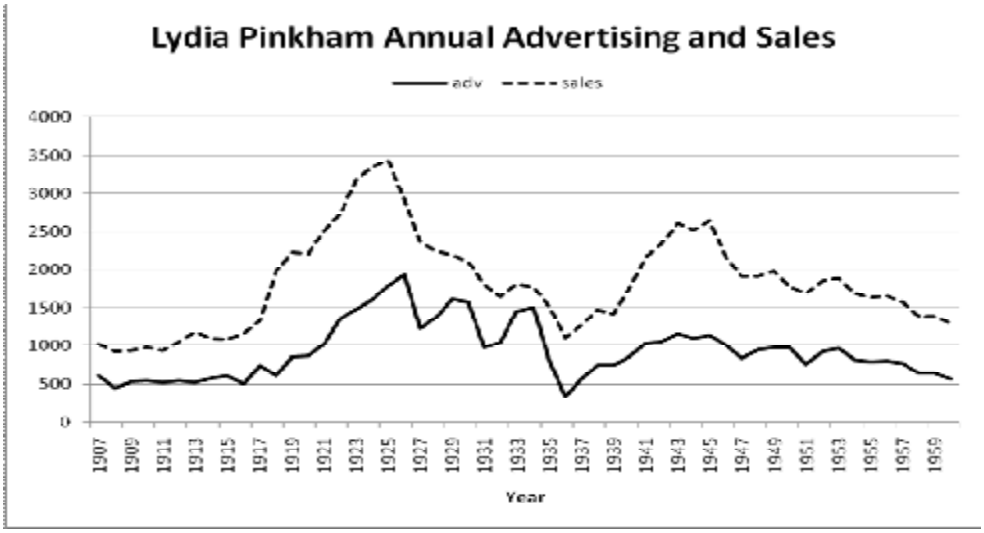


Figure 2 Time series plot for the Advertising and Sales data.

Using the heuristic method, the CPC-eigenvalues of the scatter matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ were arranged according to the sequence in $\boldsymbol{S}_1$. The eigenvalues of series 2 are descending up to $d = 5$, so r can be chosen in the interval $1 \le r \le 5$.

r	Eigenvalues for time series 1	Eigenvalues for time series 2
1	**30519735.66**	**94225252.53**
2	**20941334.05**	**71111830.77**
3	**5064088.50**	**11134196.86**
4	**3772917.66**	**2179336.42**
5	**3457664.59**	**1447251.44**
6	2573748.36	5240595.10
7	1868804.60	2588007.08
8	1037291.96	846631.05
9	487195.37	267719.61
10	401720.08	182911.91
11	275363.64	312517.49
12	219246.44	327039.47

Table 3 Comparing the eigenvalues of series 2 sorted according to series 1.

The following scree plot of the log-eigenvalues of series 2 illustrates this. The eigenvalues of S_2 are ordered according to the sequence of the first series with dots and according to its own sequence as the line.

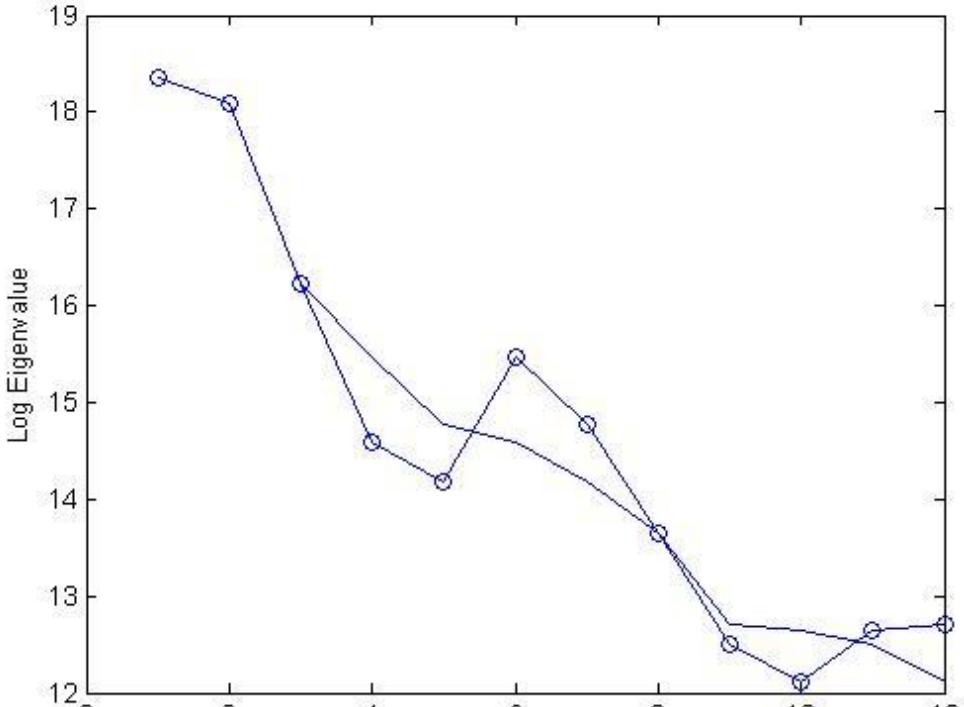


Figure 2: Series 2 log eigenvalues sorted to series 1 (dots) otherwise (line)

The hierarchical approach yielded the following table:

$r+1$	**Time series**	$\chi^2_{r+1/r}$	(df)	$\chi^2_{r+1/r}/df$
1	$\mathrm{CSSA}(1)$ vs unrelated	3.95	11	0.36
2	$\mathrm{CSSA}(2)$ vs $\mathrm{CSSA}(1)$	1.97	10	0.20
3	$\mathrm{CSSA}(3)$ vs $\mathrm{CSSA}(2)$	0.85	9	0.09
4	$\mathrm{CSSA}(4)$ **vs** $\mathrm{CSSA}(3)$	**6.13**	**8**	**0.77**
5	$\mathrm{CSSA}(5)$ vs $\mathrm{CSSA}(4)$	3.57	7	0.51
6	$\mathrm{CSSA}(6)$ vs $\mathrm{CSSA}(5)$	0.86	6	0.14
7	$\mathrm{CSSA}(7)$ vs $\mathrm{CSSA}(6)$	0.76	5	0.15
8	$\mathrm{CSSA}(8)$ vs $\mathrm{CSSA}(7)$	1.36	4	0.34
9	$\mathrm{CSSA}(9)$ vs $\mathrm{CSSA}(8)$	0.12	3	0.04

10	CSSA(10) vs CSSA(9)	0.66	2	0.33
--	--	--	--	--

Table 4 Comparing CSSA models for Advertising and Sales.

Note that at $r=3$ the improvement in the χ^2- statistic's value ($\chi^2_{4/3}$) is the highest and $\chi^2_{4/3}/8=0.77$ closest to 1 of all *r*'s less than $\tau-1$. The following figures illustrate the fit of the CSSA(3) model to both series. For Advertising: MSE =19334.09 and MAPE=10.81 and for Sales: MSE= 16266.91 and MAPE=6.03.

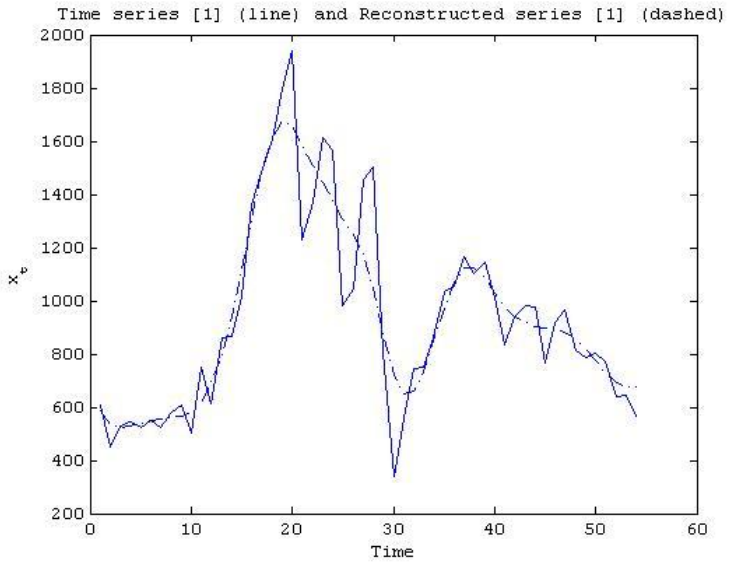


Figure 3 Advertising with the CSSA(3) fitted model.

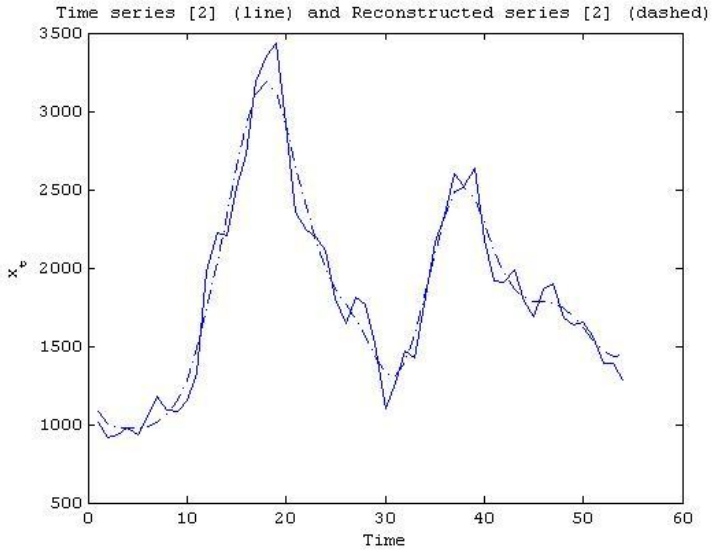


Figure 4 Sales with the fitted CSSA(3) model

These series are quite short and the CSSA fit is rather rough. The following example gives a CSSA fit to two longer economic time series.

Example 2

The Business Cycle index (BCI) for South Africa and New car sales index (NCI) is studied for the period January 1960 till December 2004. Both series are linearly increasing and co-integrated according to the Johansen test. (Johansen (1988). According to the Granger test, BCI is the leading indicator. BCI was chosen as indicated as series 1 and NCI as series 2.

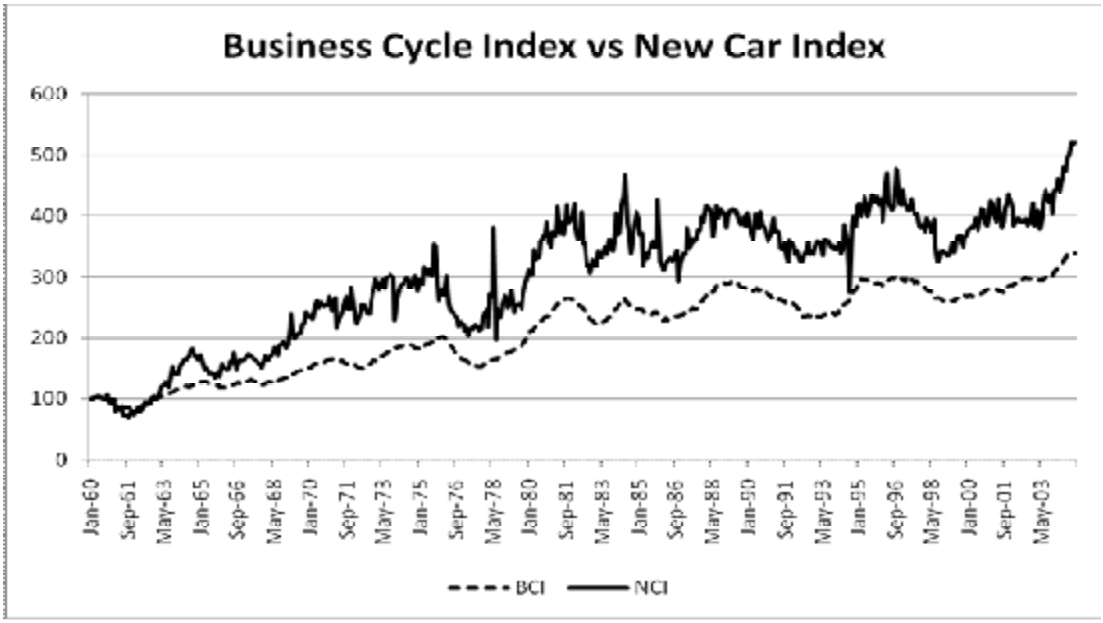


Figure 5 Time series plot for the Business Cycle Index and New Car Sales Index

From the heuristic method, the CPC-eigenvalues of the scatter matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ were arranged according to the sequence in $\boldsymbol{S}_1$. The eigenvalues of series 2 are then descending up to $d = 8$, so r can be chosen as $1 \le r \le 8$.

r	Eigenvalues for time series 1	Eigenvalues for time series 2
1	**86869413.07**	**186701030.01**
2	**3349072.75**	**9423654.37**
3	**1171954.24**	**4297880.54**
4	**300583.09**	**1488426.46**
5	**149693.55**	**674539.52**
6	**53973.58**	**304759.23**
7	**27800.64**	**225962.43**
8	**10368.25**	**199305.28**
9	8251.51	292722.28
10	8078.78	293498.28
11	4439.45	130724.62
12	4171.39	185790.49
13	3907.47	271004.90
14	2111.98	110564.06
15	1979.00	259989.93

Table 5 Comparing the eigenvalues of series 2 sorted according to series 1.

The following scree plot of the log-eigenvalues illustrates these differences among the eigenvalues of $\boldsymbol{S}_2$ when ordered according to the sequence of the first series with dots and when ordered according to its own sequence with a line.

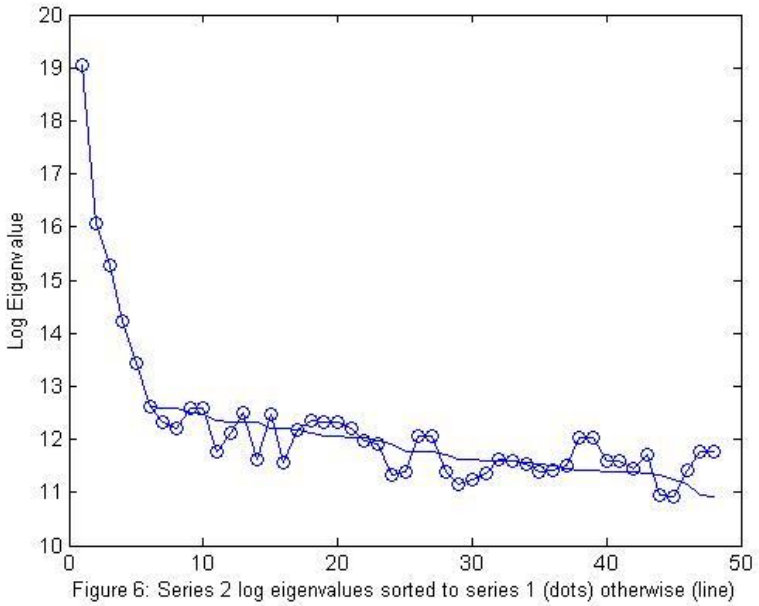


Figure 6: Series 2 log eigenvalues sorted to series 1 (dots) otherwise (line)

Figure 6 illustrates the descending order among the eigenvalues of series 2 for the different orderings and indicates a 'fit' of the CSSA(r) model for a value of r of about 7.

The hierarchical approach yields:

$r+1$	**Time series**	$\chi^2_{r+1/r}$	(df)	$\chi^2_{r+1/r}/df$
1	$\mathrm{CSSA}(1)$ vs unrelated	8.26	47	0.18
2	$\mathrm{CSSA}(2)$ vs $\mathrm{CSSA}(1)$	3.67	46	0.08
3	$\mathrm{CSSA}(3)$ vs $\mathrm{CSSA}(2)$	6.21	45	0.14
4	$\mathrm{CSSA}(4)$ vs $\mathrm{CSSA}(3)$	2.66	44	0.06
5	$\mathrm{CSSA}(5)$ vs $\mathrm{CSSA}(4)$	3.54	43	0.08
6	$\mathrm{CSSA}(6)$ vs $\mathrm{CSSA}(5)$	8.40	42	0.20
7	$\mathrm{CSSA}(7)$ vs $\mathrm{CSSA}(6)$	8.99	41	0.22
8	$\mathrm{CSSA}(8)$ **vs** $\mathrm{CSSA}(7)$	**45.61**	**40**	**1.14**
9	$\mathrm{CSSA}(9)$ vs $\mathrm{CSSA}(8)$	9.02	39	0.23
10	$\mathrm{CSSA}(10)$ vs $\mathrm{CSSA}(9)$	11.93	38	0.31
--	--	--	--	--

Table 6 Comparing CSSA models for BCI and NCS.

Note that at $r=7$ the improvement in the χ^2- statistic's value ($\chi^2_{8/7}$) is the highest and $\chi^2_{8/7}/40=1.14$ closest to 1 of all r's less than $\tau-1$. The following figures illustrate the fit of the CSSA(7) model fits for both time series.

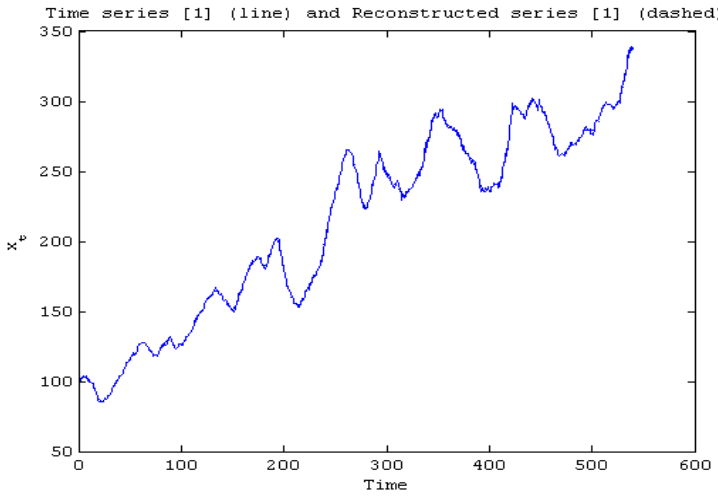


Figure 7 The Business Cycle Index with the CSSA(7) fitted model.

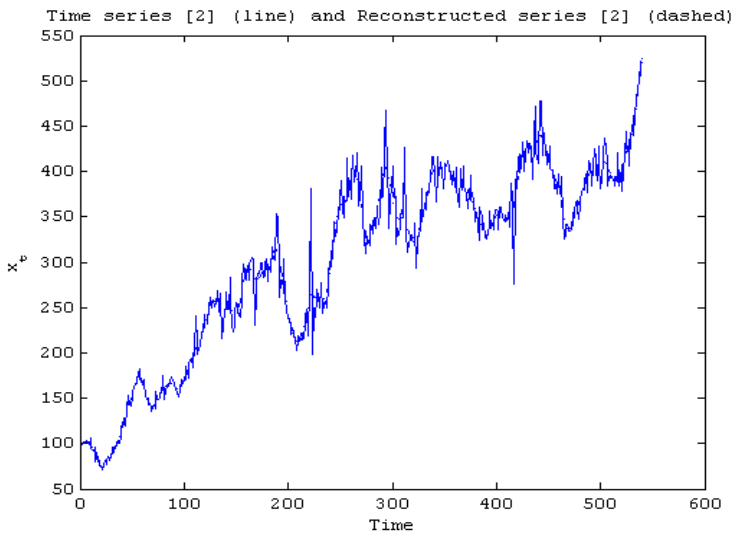


Figure 8 The New Car Sales Index with the fitted CSSA(7) model

The MSE = 2.99 and MAPE=0.58 for the Business Cycle Index and the MSE=226.20 and MAPE=3.27 for the New Car Sales Index.

7 Co-integration and CSSA.

Two time series of length $T = 100$ were simulated to investigate common properties with CSSA and co-integration:

$$y_1(t) = A_1 \cos \omega_1 t + B_1 \sin \omega_1 t + c_1 + d_1 t + e_1 t^2 + \varepsilon_1$$
$$y_2(t) = A_2 \cos \omega_2 t + B_2 \sin \omega_2 t + c_2 + d_2 t + e_2 t^2 + \varepsilon_2$$

where A_i and B_i are the amplitudes c_i, d_i and e_i, parameters for a polynomial model and ε_i a random value drawn from a normal distribution with mean 0 and variance 1 multiplied by a constant amount n_i to enlarge the noise level for each time series $y_i(t)$, where $\omega_i = 2\pi f_i$ and $f_i = 1/P_i$ where P_i is the periodicity of each time series for $i = 1,2$.

Different values for A_i and B_i, c_i, d_i and e_i where chosen and the noise levels n_i were also increased to reflect differences between the series.

In the investigation process the scree plot of the log eigenvalues of series 2 sorted to series 1 is considered. The chosen value of r $\left(0 < r \leq d\right)$, among the number d of descending singular values is then reported. If the scree plot indicate common decreasing log eigenvalues as in the examples given in section 6, Figures 2 and 6 with only minute differences among these eigenvalues of series 2, it will indicate "**fit**" of the CSSA model meaning that common features exist . In Figure 1 (see section 6) the scree plot indicates a "**non-fit**", meaning that no common features exist. In Table 2 the hierarchical method indicates a "**non-fit**" characterized by immediate high chi-square values and ratios $\chi^2_{r+1/r} \big/ df$, close to 1.

To avoid the further use of too many scree plots or tables, we will simply refer to a fit or a non-fit of the CSSA model. In the following section several scenarios are investigated to access the fit of the CSSA model. The scree plots of the log eigenvalues are investigated like mentioned above and the mean squared errors and mean absolute percentage errors of the fitted series are calculated and reported as MSE_1, MSE_2 and $MAPE_1$, $MAPE_2$ respectively for each scenario.

6.1 Investigating the window length for different values of τ :

The values for the slopes d_1 and d_2 where chosen as 0.5 while the following values where chosen for the intercepts and polynomial terms $c_1 = c_2 = 10$, $e_1 = e_2 = 0$, the amplitudes $A_1 = B_1 = A_2 = B_2 = 5$ and the period of the time series P_i =12. The noise constants where chosen as 0.5. The two time series are therefore positive linear and sinusoidal. When using $\tau = 12$ it was found that the window length is too short to detect the similarity between the two series and consequently the window length was extended to $\tau = 48$. The number of mutually decreasing eigenvalues was determined as $d = 5$ which yielded scree plots indicating fit. Consequently $r = 4$ is chosen for

the CSSA reconstruction. The reconstructed time series now fit the original time series with respective error measures MSE_1 =0.16, $MAPE_1$ =1.30 and MSE_2=0.20, $MAPE_2$=1.27. Since a window length of $\tau = 48$ gives a better fit, this window length is chosen for all further comparisons.

6.2 Influence of different slopes: Different slopes involve choosing different values of d_1 and d_2. The starting values given in (a) below are $d_1 = 0.5$ and $d_2 = 0.6$ while the other components are chosen as:

$$c_1 = c_2 = 10,\ e_1 = e_2 = 0,\ A_1 = B_1 = A_2 = B_2 = 5 \text{ and } \omega_1 = \omega_2 = \frac{\pi}{12}.$$

The noise constants were chosen as 0.5. In (a) the two slopes are almost similar, (b) gives the results when the slope of the one time series is linear positive and the second one linear negative. In (c) the slope of the first time series is linear positive but the second one is stationary. Note that in this case the series are not co-integrated.

(a) The slopes for the two time series are $d_1 = 0.5$ and $d_2 = 0.6$. The number of similarly ordered eigenvalues from CSSA is determined as $d = 4$. The scree plots indicate fit and using $r = 4$ resulted in MSE_1=0.16, $MAPE_1$=1.12, MSE_2=0.21 and $MAPE_2$=1.24. Using the hierarchical approach, $r = 4$.

(b) The slopes were chosen as $d_1 = 0.5$ and $d_2 = -2$. The number of ordered eigenvalues from CSSA is $d = 6$. The scree plots indicate fit and the use of $r = 4$ resulted in fits with MSE_1 =0.18, $MAPE_1$=1.29, MSE_2=0.25 and $MAPE_2$=0.98. The hierarchical approach yielded, $r = 4$.

(c) The slopes where chosen as $d_1 = 0.5$ and $d_2 = 0$. The number of ordered eigenvalues is $d = 1$ which immediately indicate non-fit of the CSSA model. This is also clear from the scree plots and by comparing the MSE's and MAPE's after the CSSA model is fitted. Using $r = 1$ results in MSE_1=24.59,

$MAPE_1$=17.12, MSE_2=24.22 and $MAPE_2$=68.82. The Hierarchical approach indicates a non-fit. Immediate high chi-square values and ratios $\chi^2_{r+1/r}/df$, close to 1 (as in Table 1) is obtained.

Note that the two series in (a) and (b) are both first order non-stationary and the CSSA model fits, indicated by the small MSE's and MAPE's. It seems that a difference in the slopes when both series are first order non-stationary does not influence the fit of the CSSA model, though it influences co-integration. However in (c) the one series is first order non-stationary while the second one is stationary. In this case the series are not co-integrated but the CSSA model also does not fit.

6.3 Influence of different quadratic components: In the following analysis (a) the values for the linear coefficients are fixed on $d_1 = d_2 = 0.5$ but the quadratic coefficients are chosen as $e_1 = 0$ and $e_2 = 0.05$. Thus the one time series has a linear trend while the other one has a quadratic trend. The noise variance is chosen as 0.5, while the other components were chosen as $c_1 = c_2 = 10$, $A_1 = B_1 = A_2 = B_2 = 5$ and $\omega_1 = \omega_2 = \frac{\pi}{12}$. In the analysis (b) the linear coefficients are changed to $d_1 = 0.5$ and $d_2 = -2$ resulting in a negative linear series 2 being also perpendicular to series 1.

(a) The quadratic components were chosen as $e_1 = 0.05$ and $e_2 = 0$. The number of ordered eigenvalues is $d = 2$. Using $r = 2$ yielded the following fitted series with CSSA: MSE_1 =13.35, $MAPE_1$=7.00, MSE_2 = 11.31 and $MAPE_2$=12.68. From the hierarchical approach, $r = 3$.

(b) The slopes were chosen as $d_1 = 0.5$, $d_2 = -2$ and the quadratic components as $e_1 = 0.05$ and $e_2 = 0$. The ordered eigenvalues $d = 3$. The Hierarchical approach gives immediate high chi-square values indicating a non-fit.

Choosing $r = 3$ both the scree plot and the error measures MSE_1=9.70, $MAPE_1$=7.20, MSE_2=86.92 and $MAPE_2$ = 42.98 also indicate no fit.

In both cases (a) and (b) the large MAPE's indicate bad fit, particularly for the second case as in (b). Note that since the one series is linear, i.e. of order I(1) and the other quadratic of order I(2), they are not co-integrated and the CSSA model does not fit.

6.4 Influence of periodicity on CSSA: The periods of the two time series, ω_1 and ω_2 were chosen as different, $\omega_1 = \frac{\pi}{12}$ and $\omega_2 = \frac{\pi}{6}$ while the other components were chosen the same, $A_1 = B_1 = A_2 = B_2 = 5$, $c_1 = c_2 = 10$, $d_1 = d_2 = 0.5$, $e_1 = e_2 = 0$ and noise constant 0.5. Three ordered eigenvalues ($d = 3$) were observed. Using $r = 3$ and fitting the CSSA model resulted in MSE_1=0.21, $MAPE_1$=1.31, MSE_2=25.50 and $MAPE_2$=18.09. The Hierarchical approach indicates a non-fit of the CSSA model due to immediate large chi-square values. This is also visible from the scree plot and larger MSE and MAPE of the second series. Since the two series are however co-integrated according to the Johansen (1991) test, this illustrates that co-integration not necessarily imply that the CSSA model will fit. Thus CSSA is a stronger condition or measure of similarity between two time series than co-integration.

6.5 Influence of different amplitudes on CSSA: Different amplitudes were given to the two time series. The time lag *k*, was set to 0 and the other components were chosen as $d_1 = d_2 = 0.5$, $e_1 = e_2 = 0$, $c_1 = c_2 = 10$ and $\omega_1 = \omega_2 = \frac{\pi}{12}$.

(a) The amplitudes were chosen as $A_1 = B_1 = 5$, $A_2 = B_2 = 7$. The ordered eigenvalues indicate $d = 6$. The scree plot indicate fit and when choosing

$r = 5$, the following error measures were found MSE_1=0.18, $MAPE_1$=1.23, MSE_2=0.22 and $MAPE_2$= 1.54. Using the hierarchical approach $r = 5$.

(b) The amplitudes were chosen as $A_1 = B_1 = 5$ and $A_2 = B_2 = 15$. The number of ordered eigenvalues is $d = 1$, which immediately indicates a non-fit, even without plotting any scree plot. If the CSSA model is fitted the error measures are MSE_1=25.03, $MAPE_1$=17.38, MSE_2=222.74 and MAPE2=224.51, confirming the bad fit. The hierarchical approach's large chi-square values also confirm the non-fit.

For ratios of amplitudes A_2/A_1 and B_2/B_1 close to one, the CSSA model gives a good fit, but when these amplitude ratios get larger, the CSSA model does not fit. In both cases (a) and (b) the two series are still co-integrated.

8. CONCLUDING

Common singular value decomposition can be used to investigate common structure among different time series. The common structure is explained in terms of the r singular values resulting from the common submatrix $\boldsymbol{B}: \tau \times r$ in the matrix/matrices $\boldsymbol{\beta}_i = \left[\boldsymbol{B} \quad \boldsymbol{B}_i^{(2)} \right] = \left[\underline{b}_1 \quad \ldots \quad \underline{b}_r \quad \underline{b}_{r+1}^{(i)} \quad \ldots \quad \underline{b}_\tau^{(i)} \right], i = 1,2$, used to simultaneously diagonalize the scatter matrices $\tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{X}}_i'$ to CPC or CPC(r) structure in the Flury-Gautschi sense, where $\tilde{\boldsymbol{X}}_i \left(i = 1,2 \right)$ are the respective centered trajectory matrices. These vectors are the base vectors of the common R-flat. Failure to simultaneously diagonalize these matrices with CPC or CPC(r) indicates that no common structure exists at all and thus no common R-flat exists.

CSSA gives more insight into the nature and dimensionality of common features among time series than co-integration alone.

REFERENCES

Broomhead, D.S. and King, G.P.,1986a. Extracting qualitative dynamics from experimental data. Physica D 20, 217-236.

Broomhead, D.S. and King, G.P., 1986b. On the qualitative analysis of experimental dynamical systems. In S. Sarkar (Ed.), Non-linear phenomena and chaos. 113-144. Adam Hilger, Bristol.

Buchstaber, V.M., 1994. Time series analysis and Grassmannians. Amer. Math. Soc. Transl 162 Series 2, 1-17.

Danilov, D.L., 1997. Principal components in time series forecast. Journal of computational and Graphical Statistics, 6, 112-121.

Elsner, J.B. and Tsonis, A.A., 1996. Singular Spectrum Analysis. A New Tool in Time Series Analysis. New York: Plenum Press.

Engle, R.F. and Granger, C.W.J., 1987. Co-integration and error correction: representation, estimation and testing. Econometrica 55, 251-276.

Flury, B., 1984. Common principal components in k groups. Journal of the American Statistical Association 79, 892-898.

Flury, B., 1988. Common principal components and related multivariate models. New York: Wiley.

Flury, B. and Gautschi, W., 1986. An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. SIAM Journal of Scientific and Statistical Computing 7, 184-196.

Golyandina, N., Nekrutkin, V.V. and Zhigljavsky, A., 2001. Analysis of Time series Structure SSA and Related Techniques. Boca Raton, Chapman & Hall/CRC.

Johansen, S., 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12, 231-254.

Johansen, S., 1991. Estimation and hypothesis testing of cointegration vector in Gaussian vector autoregressive models. Econometrica 59, 1551-1580.

Vautard, R., Yiou, P. and Ghil, M. 1992. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. Physica D, 58, 95-126.

Venter, J.H., 1998. Forecasting by identification of linear structure in a time series. Unpublished talk presented at the 1998 Conference of The South-Africa Statistical Association.

Viljoen, H and Nel D.G., 2009. Common singular spectrum analysis of several time series. Journal of Statistical Planning and Inference (In press)

Wei, W.W.S., 2006. Time Series Analysis: Univariate and Multivariate Methods. (Sec. Ed.) Boston: Pearson Addison Wesley.