In the Name of Allah

The Sixth International Statistics Conference

Full Papers

August 26-28, 2002

Tarbiat Modares University

Tehran, Iran

PREFACE

This volume contains the abstracts of invited and contributed papers presented at the Sixth International Statistics Conference (ISC6). This conference is a three day conference held every two years with cooperation the Iranian Statistical Association. ISC6 is jointly organized by Department of Statistics and Department of Biostatistics of Tarbiat Modares University. It is taking place from 26 to 28 August at Tarbiat Modares University, Tehran, Iran. Previous ISCs from 1992-2000 were at Isfahan University of Technology, Ferdowsi University of Mashad, Iranian Statistical Center, Shahid Beheshty University, Isfahan University of Technology respectively.

The scientific committee of the conference selected 193 abstracts for oral presentations, 73 abstract for posters, and 5 abstracts for workshops. from over 340 submitted abstracts.

We would like to thank our colleagues and all who helped us for this publication. We gratefully acknowledge the careful efforts of R. Safari, A. Kavoosi, S. Jolani and S. Khafry to made the publication of abstracts possible.

M. Mohammadzadeh

August 2002

Organizing Committee

Mohammadzadeh, M., (Secretary), Tarbiat Modares University

Vahidi-Asl, M. Q. (Chair), Shahid Beheshti University

Kazemnejad, A., Tarbiat Modares University

Hosseini, S. M., Tarbiat Modares University

Faghihzadeh, S., Tarbiat Modares University

Babaei, G. R., Tarbiat Modares University

Scientific Committee

Vahidi-Asl, M. Q. (Chair), Shahid Beheshti University

Meshkani, M. R., Shahid Beheshti University

Parsian, A., Isfahan University of Technology

Mohammadzadeh, M., Tarbiat Modares University

Kazemnejad, A., Tarbiat Modares University

Faghihzadeh, S., Tarbiat Modares University

Navabpoor, H. R., Allameh University

Grami, A., Statistical Research Center

Mehrabi, Y., Shahid Beheshti University of Medical Science

Table of Contents

ORAL PRESENTATIONS

Convergence of Weighted Sums Of r.v.s Using Sub - Gaussian Techniques

Amini, M. and Bozorgnia, A.

A11098

Department of Statistics, Ferdowsi University, Iran.

Abstract. In this paper, we study some strong limit theorems for the sequence $\{1/n^{\beta}\sum_{i=1}^{n} X_n, n \geq 1\}$ for each $\beta > 0$ and weighted sum $\sum a_{nk}X_k$ where $\{X_{n,,} \geq 1\}$ is a sequence of negatively dependence sub - Gaussian random variables and a_{nk} is an array of nonnegatively real numbers.

Keywords: Negatively Dependent Random Variables, Strong Law of Large Numbers, Weighted Sums.

1 Introduction

Convergence theorems for weighted sums have been obtained by Chow [] Hamson[3].
Pruitt [5] ,Bozorgnia et . al. [2] , Amini et.al. [1] ,and for independent,generalized Gaussian r.v.by Chow [3],and by Taylor and Chung Hu [6].

Lemma 2 (a)Let $X_1, ..., X_n$ be ND random variables and $f_1, ..., f_n$ be a sequence of Borel-fanctions which are monotone,then $f_1(X_1), ..., f_n(X_n)$ are ND random variables.
Let $X_1, ..., X_n$ be ND random variables, then

$$E(\prod_{1}^{n} X_j) \leq \prod_{1}^{n} E(X_j)$$

Definition 2. A Symetric r.v is said to sub-Gaussian r.v. if these exists a real number $\alpha \geq 0$ such that for each real number t

$$Ee^{tx} \leq exp[\frac{\alpha^2 t^2}{2}]$$

lemma 1. If X is a sub-Gaussian r.v. with

$$\tau(X) = inf\{\alpha \geq 0 : Ee^{tX} \leq exp[frac\alpha^2 t^2 2]\}$$

then

a)$E[e^{t(X)}] \leq 2exp[\frac{\alpha^2 t^2}{2}] \quad , t \in R$

b) $P(|X| \, \varepsilon) \leq 2exp[\frac{-\varepsilon^2}{2\alpha^2}], \quad \varepsilon > 0$

c)If $|X| \leq M$ then $\tau(X) \leq \sqrt{2}M$.

Theorem 1 .Let $\{X_n, n \geq 1\}$ be a sequence of stochastically bounded by random variables X , with $E(X^2)$ and let $\{a_{nj}, 1 \leq j \leq n\}$ be a triangular array of real numbers with $\sum_{j=1}^{n} a_{nj}^2 = O(n^{-\beta}), \beta > 1$, then

$$\sum_{1}^{n} a_n j X_j \to 0$$

Theorem 2 . Let $\{X_n, n \geq 1\}$ be a sequence of ND random variables with $EX_j = 0$,

a)if $sup|X_n| \leq C$ and $\max|a_{nj}| = O(n-1)$,then

$$\sum a_{nj} X_j \to 0 \ hspace1cma.e$$

b)If $|X_n| \leq C, a.e$ and $\max|a_{nj}| = O(n^{-1}), \beta > 1/2$ then

$$\sum a_{nj} X_j \to 0 \ hspace1cma.e$$

c)If $B_n^2 = \sum_1^n \sigma_j^2$ and $\sum exp\{-\frac{2\varepsilon^2}{\alpha\beta_n^2 \sum a_{nj}^2}\}$ then

$$\sum a_{nj} X_j \to 0 \qquad a.e$$

Theorem 3. Let $\{X_n, n \geq 1\}$ be a sequence of ND sub-Gaussian r.v.s, with $\tau(X_n) \leq \alpha_n$

i)$S_n = \sum_1^n X_k$ is a sub-Gaussianr.v. with $\alpha^2 = \sum_1^n \alpha_i^2$.

ii) If $\sum_1^n \alpha_i^2 = O(n^{2-\beta}), \beta > 0$,then

$$\lim_n \frac{1}{n} \sum_{1}^{n} X_k \qquad a.e$$

iii) If $\alpha_1 = \alpha_2 = ... = \alpha_n = \alpha$ then for some $\beta > \frac{1}{2}$,

$$\lim_n \frac{1}{n_\beta} \sum_{1}^{n} X_k \qquad a.e$$

Teorem 4 . (a) Let $\{X_n, n \geq 1\}$ be a sequence of ND random variables satisfying $P(a \leq X_i \leq b) = 1$, then

$$\lim_n \frac{1}{n^\beta} \sum_{1}^{n} (X_k - E(X_k)) = 0 \qquad a.e$$

b)If $\{X_n, n \geq 1\}$ is ND and identically distriduted r.v.s with $E(X_1) = 0$, $var(X1) = 1$, $E(X_1^k) < \infty$ then $\frac{S_n}{\sqrt{n}}$ is an asymptotically sub-gaussian r.v.

teorem 5 .Let $\{X_n, n \geq 1\}$ be a sequence of ND sub-gaussian r.v.s

i)If $\lim_n \sum_1^n a_{nk}^2 = l \neq 0 < \infty$, then for $\beta > 0$,

$$\lim n^{-\beta} \sum a_{nk} X_k = 0 \qquad a.e$$

ii)If $a_{nk} = O(n^{-\beta})$ for some $k \leq n$ and $\beta > \frac{1}{2}$,then

$$\lim_n \sum a_{nk} X_k = 0 \qquad a.e$$

iii)If $\sum a_{nk}^2 = O(n^{-\beta})$, $\beta > 0$,then

$$\lim_n \sum a_{nk} X_k = 0 \qquad a.e$$

2 Some strong limit theorems for weighted sums

In this section we some obtain some strong limit theorems for weithed sums .let $T_n = \sum_{k=1}^{\infty} a_{nk} X_k$ and $T_{nk} = \sum_{k=1}^{\infty} a_{nk} X_k$ when $\{X_n, n \geq 0\}$ is a sequence of negetivly dependence sub -Gaussian r.v.s and $\{a_{nk}\}$ is an array of real numbers.
We prove $T_n = \sum_{k=1}^{\infty} a_{nk} X_k$ converges a.e underr the condition that $E(X_n|F_{n-1}) = 0$ where $f_n = \sigma(X_1, ..., X_n)$ and $\sum_k^{\infty} \sigma_{nj}^2 = O(K^{-\beta}), \beta\, 0$

lemma 2 . Let $\{X_n, n \geq 1\}$ be a sequance of ND sub-Gaussian r.v.s with $\tau(X_k) \leq \alpha$, then

i)T_n is a sub-Gaussian r.v . with $\tau(T_n) \leq \alpha\sqrt{a_n}$

ii) for every $\varepsilon\, 0$

$$P(|T_n|\, \varepsilon) \leq 2exp[-\frac{\varepsilon^2}{2\alpha^2 A_n}]$$

were $A_n = \sum_{k=1}^{\infty} a_{nk}^2$.

Corollary 1. If $\sum 2exp[-\frac{\varepsilon^2}{2\alpha^2 A_n} \leq \infty$,then

$$\lim_{n\to\infty} suma_{nk} X_k = 0 \qquad , a.e$$

In particular if $A_n = O(ln^{-1}(n))$ then (1) holds.

ii) If$S_n = \sum X_k$ and $\beta > 0$, then

$$\lim n^{1/2} (ln^{-(1+\beta)/2}(n)) S_n = 0 \qquad a.e$$

Teorem 6 (a) Let $\{X_n, n \geq 1\}$ be a sequance of ND sub-Gaussian r.v.s then for every $x \in R$

$$P(max|T_{nj}| > x) \leq 2exp[-\frac{x^2}{2\alpha^2 A_n}]$$

b)If $\{T_n m, m \geq 1\}$ converges in probability for every n , then it converges a.e.

c) $T_n = \sum a_{nk} X_k$ convereges a.e . for each n .

References

Amini,M .And Bozorgnia , A . *Negativiy Dependebt Bounded r.v.s.* .Jornal of Applied Mathematics Analysis 13:3 (2000)261-267.

Bozorgnia,S .Patterson , R.F and Taylor R.L *Limit Theorems for Dependent r.v.s.* World Congress Nonlinear Analysis 92,1996,1639-1650

Chow,Y.S *Some Convergence Theorem for Independent r.v.s Ann.* Math. statist. 37,1966,1482-1493.

Hanson ,D.L. and Koopman L.H.*On the Convergence Rate of the Law of Large Numbers for Linear Conslinations of Independent r.v.s.*Ann.Math. stat. 36,1965,559.

Pruitt ,W.E.*Summability of Independent r.v.s.* J. Math. Mech. 15, 1966, 769-776.

Taylor, R.L. *Complete Convergence for Weighted Sums fo Array of Elements.* J. Math. Science Vol .6,1983. no . 1,P.69-79

The Effect of Different Parameters on Students Scores

Badshah, S.

P17023

Islamia College, University of Peshawar, Pakistan.

Abstract. This study was designed to estimate the effect of different parameters (i.e. SSC marks, mathematics, statistics, handwriting, attendance, hostel, study hours, parent's education and medium of schools) on student marks in FA Intermediate part-I. The effects of mathematics, statistics, handwriting, attendance and educated mother are significantly positive whereas those of "English Medium" schools are significantly negative. Father's education, study hours and position among siblings have no significant effect on one's academic achievement. This study also shows that weak students are more studious.

1 Introduction

There are verbal debates among parents of the students about different parameters effecting students' marks at the time of admission in colleges and their results. The common parameters, which the authors considered are SSC marks, mathematics, statistics, handwriting, attendance, hostel, study hours, parent's education and medium of school. Some people prefer hostel for studies whereas a few do not do so. Some teachers advocate that students having a sound background in mathematics and statistics get higher marks. Some teachers insist on improving the handwriting of students because it has a positive effect on evaluation of papers. According to some views about "English Medium Schools", their students are bold, take part in extra-curricular activities speak better English but they are generally weak in mathematics. Attendance also plays a very important role in a student's achievement. To study the above parameters, their effects and importance, a study is designed based on 112-second year students of Islamia College Peshawar. The methodology and results is given below.

2 Material and Methods

To study different parameters related to one's marks in intermediate arts, part-I, the data of 112 students, sessions 1998-99 and 1999-2000, are collected through a questionnaire*. The parameters studied are: First year marks in intermediate (FM), Marks in Mathematics (MM), Marks in Statistics (MS), Boarder or day-scholar (H), Father's Education (FE), Mother's Education (ME), Position among siblings (PB), Study hours (SH), Handwriting (WR)

"fair or not", Attendance percentage (AP), Medium of School in SSC (SC) and Marks in matriculation (SSC). The Intermediate part-I marks are taken as dependent variable and the remaining parameters are independent variables. The effects of different parameters on 112 students were studied. The details are given below:

The data shows that the statistic for Intermediate part-I marks recorded are 112 (with zero missing values.), having mean marks 325.01, with minimum marks of 229 and maximum of 409. The mean of the distribution is 610.95 in S.S.C with minimum marks of 529 and maximum of 651 etc. as shown in table 1.

Parametes	Median	Minimum	Maximum
Intermediate Part-I Marks	332	229	409
hline S.S.C Marks	612	529	651
Marks in Mathematics	70	33	97
Marks in Statistics	65.5	37	98
Attendance in Study hours	4	1	8

Among these students 56.3% were found boarders and 39.3% day-scholars with 4.4% missing. In father's education category, out of which 65.2% were found literate whereas 32.1% illiterate, while 2.7% missed the said parameter. 110 recorded their mother's education, where it is found that 21.4% are literate,76.8% are illiterate, and 1.8% missed this column. 41.1% are from English medium, and 58% are from Urdu medium schools. 0.9% missed this column. The handwriting of 69.6% students is found fair, 27.7% is not fair and 2.7% missed this category. It is found that 27.7% are the first among siblings who joined the college.

Linear Model:

Stepwise selection Criteria was used in the selection of the model, i.e. "Probability-of-F-to-enter ≤ 0.05 , Probability-of-F-to-remove ≥ 0.10". The models estimated are given below:

$$FM = f(MM, MS, BD, FE, ME, PB, SH, WR, AP, MS) \qquad (1)$$

The above model is not only effected by quantitative variables (MM, MS, SH and AP), but also qualitative variables to capture the effect of the qualitative variables (BD, FE, ME, PB, WR and MS).

The Variance Inflation Factor (VIF) for every regression coefficient was calculated, to detect the multicollinearity (John, Neter,. 1987, pp: 382-92)

effects on the regression coefficients and models themselves. " A maximum $(VIF)_k$ in excess of 10 is often taken as an indicator that multicollinearity may be unduly influencing the least squares equation." (John, Neter., 1987, pp: 391-3).

Effect of Parameters:

To estimate the regression coefficients of all variables (Quantitative and Qualitative) in the model, all variables were entered in the model simultaneously as given below:

$$FM = \beta_0 + sum_{i=1}^{11}\beta_i P_i$$

where P_i represents the different parameters under study. That is

$$FM = \beta_0 + \beta_1(MM) + \beta_2(MS)\beta_3(AP) + \beta_4(WR) + \beta_5(H) + \beta_6(SSC) + \beta_7(SC) + \beta_8(ME) + \beta_9(FE) + \beta_10(PB) + \beta_11(SH) \quad (2)$$

Where

BD = "1" for Boarder and "0" otherwise.
FE = "1" for Educated Father, "0" otherwise.
ME = "1" for Educated Mother, "0" otherwise.
PB = "1" for First in brothers, "0" otherwise.
WR = "1" for Fair writing, "0" otherwise.
SC = "1" for English Medium, "0" otherwise.

The estimated regression coefficients in model (1) are given below:

$$\begin{aligned} FM = {} & -63.90 + \underset{(6.61)^*}{1.01(MM)} + \underset{(6.31)^*}{1.11(MS)} + \underset{(5.08)^*}{0.92(AP)} + \underset{(3.01)^*}{14.02(WR)} + \underset{(2.15)^{**}}{9.53(H)} \\ & + \underset{(2.84)^*}{0.26(SSC)} - \underset{(-2.38)^{**}}{9.88(SC)} + \underset{(1.89)^{**}}{9.64(ME)} + \underset{(0.34)^{***}}{1.48(FE)} + \underset{(0.23)^{***}}{0.98(PB)} - \underset{(-0.54)^{***}}{0.65(SH)} \end{aligned} \quad (3)$$

(Figures in parenthesis are t-ratios)
$R - square = 0.79, F = 30.4^*$and
$1.04 \leq (VIF) \leq 1.60$
for all coefficients in the above*model*.

And (*) for "Sig. at 1%", (**) for "sig. at 5%" and (***) for "not sig. at 5%", which shows that FE, PB and SH are not significant at5% where as H, SC, and ME are significant at5%. SC and SH shows negative effect. The remaining variables are significant at 1%. (David, R., 1991, pp: 281-4, 89) When Stepwise selection Criteria (John, Neter.,1987, pp: 430-5), was used in the selection of the model: i.e. "Probability-of-F-to-enter *leq* .050, Probability-of-F-to- remove $\geq$ .100". The model (2) and (3) emerged as given

below with their R-squares, F- ratio and t-statistic (David, R., 1991, pp: 401-4).

$$FM = \underset{}{-58.14} + \underset{(6.91)^*}{1.03(MM)} + \underset{(6.61)^*}{1.08(MS)} + \underset{(5.19)^*}{0.92(AP)} + \underset{(3.12)^*}{14.29(WR)} + \underset{(2.16)^{**}}{8.99(H)}$$

$$+ \underset{(2.82)^*}{0.25(SSC)} - \underset{(-2.37)^{**}}{9.48(SC)} + \underset{(2.13)^{**}}{10.15(ME)} \quad (4)$$

$R - square = 0.79, F = 42.97*$
$1.20 \leq (VIF) \leq 1.50$
for all coefficients in the above model.
(Figures in parenthesis are t-ratios)

And (*) for "Sig. at 1%", (**) for "sig. at5%" and (***) for "not sig. at5%". Among all these models, model (3) is more appropriate, on the bases of its F=68.34, which is the highest among the models selected. Which shows that the model 3 and its regression coefficients are highly significant at 1

$$FM = \underset{}{86.60} + \underset{(5.08)^*}{1.22(MM)} + \underset{(7.95)^*}{1.09(MS)} + \underset{(6.18)^*}{0.84(AP)} + \underset{(4.61)^*}{17.91(WR)} \quad (5)$$
$$\underset{(3.74)^*}{}$$

$R - square = 0.74, F = 68.34^*$and
$1.10 \leq (VIF) \leq 1.40$
for all coefficients in the above model.
(Figures in parenthesis are t-ratios)

(*) For "Significant at 1 %".

3 Conclusions

1. This study reveals that Marks in Mathematics, Marks in Statistics, Attendance, and Handwriting are significantly positive at $\alpha = 0.01$.
2. S.S.C marks, Hostel stay and Mother's education has significantly positive effects at $\alpha = 0.05$ on students performance in Inter part-I examination.
3. Equation 2 shows that the performances of English Medium schools are weaker as compared to Urdu medium schools (the coefficient of "SC" in equation 2 is negative).
4. Father's education, Position among siblings and study hours have no significant effect.
5. The Coefficient of study hours "SH" is negative in equation 2, which shows that relatively dull students are more studious.

References

David, R., (1991), *Introduction to Statistics*, 2nd Edition, West Publishing Company, St. Paul New York, pp: 281-4, 289, 401-04.

Koutsoyiannis, A., (1988), *Theory of Econometrics*, 2nd Edition, University of Ottawa, pp: 117-36.

John, Neter., (1987), *Applied Statistical models*, 2nd Edition, Irwin, Inc. Homewood, Illinois 60430 Toppan Company Ltd. Tokyo, pp: 382-92, 391-3, 430-5.

∗ The Questionnaire used:

1) Marks in SSC————————, 2) Marks in Mathematics————,

3) Marks in Statistics————————, 4) Study hours————————,

5) Attendance in percentage————, 6) Boarder ———————— (Yes/No),

7) Handwriting is fair (Yes/No), ——-8) Father is educated ——— (Yes/No),

9) Mother is educated (Yes/No),———10) Medium of school (English/ Urdu),

11) First among siblings (Yes/No).

Inflattionary Trends in Marks and Quality of Education

Badshah, S.

P27023

Islamia College, University of Peshawar, Pakistan.

Abstract. The average marks of the students are increasing day by day, where as the relative quality of education is decreasing accordingly. A new model of question paper has been devised after careful study of the existing question papers. The proposed model question paper will not only reduce the quantity of marks obtained, selective teaching and selective studies but it will also improve attendance of the students, quality of education, teachers and student-teacher relationship. It will also help in reducing the non-productive activities in educational institutions.

1 Introduction

A technical and rational administrative decision has to be based on database, which not only solves the direct problems but also plays a very important role in solving indirect problems as well. Inflationary Trend in examination marks started a few years ago, with the result that now every student scores such a high percentage of marks, that even the parents do not accept and express their opinion by saying that "Standard of education is dropping day by day".

Prior to 1980, only 1st division was sufficient for admission to professional college, where as from 1994 and onwards, every normal student in NWFP gets 800 plus marks out of 1100 at Intermediate (Science) level. The case was different in other provinces of the country and the inflationary trend of marks was present even before 1994. The students of NWFP always criticized such abnormally high marks. In 1995, the students, as well as their parents protested against the high percentage of marks in examinations of other province for admission in Khyber Medical College NWFP Pakistan. As a result the Govt. of NWFP decided in 1996 to conduct "Entry Test" for the admission to Medical and Engineering Colleges to reduce the effect of inflationary trend in marks. This in turn has created a problem for the students and their parents. Now the students not only have to get high marks in Intermediate (Science) examinations, but also have to go through the coaching classes for "Entry Test". Most of the centers for the coaching classes are located in urban centers of the province. As such participation in the classes has become easy for students living in urban areas, but not for poor students living in rural areas. Therefore, the candidates of urban areas are able to get more marks and the poor students are badly effected. As such the candidates from rural areas are deprived from admission to professional colleges.

The race of inflationary marks is primarily because of the style of question papers, which allows 50% of the course to be enough for solving %100 of the question paper (i.e., 5 out of 10 questions). As a result

This creates a lot of problems for teachers in the class, because the students miss about%50 of the classes, and the teachers' interest also decreases. Therefore, the student-teacher interaction decreases.
The students residing in hostels lose interest in studies and get involved in creating problems for administration.
They cover %50 of the course in one month through private tuition, and are not serious in attending classes. Therefore, the quality of education is lower whereas percentage of marks obtained increases.
The students have more opportunities to get involved in politics, because they can cover their courses in half the time (%50 course) and rest of the time is used for such activities.

Therefore, it is important to have a standard format of question papers for all subjects having at least the following properties:

1. Compel students to study %100 courses and work hard, and thus compel students to attend maximum classes.
2. The question paper should enhance the award of standard marks and reduce the inflationary trend in marks.
3. The question paper should test the real knowledge of the students.
4. It should be in a style, which discourages private tuition, and encourages class- room teaching.

2 Analytical Investigation

The above mentioned problem of inflationary trend in marks has been studied through careful observation of the various question papers from B.I.S.E. and the universities as described in the following sections.

2.1 Material and Methods

The style of a question paper and instructions for attempting the question papers at B.I.S.E. (for Intermediate) and University of Peshawar (for Bachelors and Masters) and Gomal university (for Masers) papers have been studied (Table- 2, 3). It is found that all the question papers are of different nature. The numbers of questions, parts of papers and instructions for attempting the questions are different in different subjects even with in the same institutions. Also the result of B.I.S.E. Peshawar for a period of 1981-94 was collected and analyzed to study the trend in Intermediate (Science) marks obtained by the students. It is found that found that there is a highly positive and significant

$F = 8.77038, P - value = 0.0119$ linear relationship with 62.07 increase in grades A and B every year. (See Table-1)

Table 1.

Variable	B
Year	62.072527*
(Constant)	726.527473**

Significant at 0.05,**significant at 0.01

Instituitions		Category	Subjects	Total Q's given	Comp .Q's	Parts	Attpt. any	Choice
B.I.S.E Peshawar	A	A1	Econometrics Statistics Civics Isl./History Philosophy	10	-	-	5	5 out of 10
		A2	Mathematics Physics Chemistry	10	one	A,B,C	2=A 2=B	6 out of 10
		A3	English Urdu Pashto	8	-	-	8	OR among parts of a Q's
		A4	Islamiat	5	-	-	5	OR among parts of a Q's
		A5	Biology	8	Tow	A,B	3 out of 6	5 out of 8
Peshawar University	B	B1	All subjects MA/M.Sc	10	-	-	5	5 out of 10
Gomal University		B2	All subjects MA/M.Sc	7	-	-	4	4 out of 7

Table 2. Style of Question Papers of different Subjects of different Institution

Source: Exam.Section B.I.S.E Peshawar,U.O.Peshawar and Gomal University D.I.Khan

2.2 Interpretation of data

2.3 Category A

Section A1 Includes Economics, Statistics, Civics, Islamiat/History and Philosophy. Total questions given in this section are 10, with no parts (such as part A, B or C), no compulsory question and with instruction to attempt any five questions out of ten (i.e.,100% choice).

Section A2 Includes Mathematics, Physics and Chemistry. Total questions given in this section are ten, with one compulsory question and the remaining nine are divided in to three parts (i.e., parts A, B and C), with instructions to attempt a total of five questions from the remaining nine questions at least one question from each section. Here the total questions to be attempted are six.

Section A3 Includes English, Urdu and Pashto. Total questions in this section are eight, having "OR" between two parts of a question. Here the total questions to be attempted are eight (thus in reality eight out of sixteen questions i.e., 100% choice).

Section A4 Includes only one subject Islamiat and is different from all the given question papers discussed above. Total questions given are five, with the instruction to attempt all questions, and there is "OR" between two part of each question (i.e., 100% choice).

Section A5 Includes only one subject of Biology. Here total number of questions is eight with instruction that first and second questions are compulsory and the remaining six questions are divided into two parts (A and B) with instruction to attempt any three questions from the remaining six questions. Thus five questions out of eight have to be attempted.

2.4 Category B

This category is divided into two sections (i.e., B1 and B2), which includes question papers of Peshawar and Gomal Universities discussed in detail as follows.

Section B1 Includes all MA/M.Sc question papers of the University of Peshawar. The question papers are of same type. Total questions given are 10, with the instructions to attempt any five questions (i.e., choice is 100%). Section B2 Includes all question papers of MA/M.Sc of

Gomal University. Total questions given are seven, with the instruction to attempt any four questions (i.e.,75% choice).

2.5 Table 3

It shows BA/B.Sc. style of the question paper of the University of Peshawar. The question papers are classified in ten categories (C1 to C10). Every category is different from the other. The major category is C3 and includes seven subjects. In this category total questions are 10, with instructions to attempt any five.

The second major category is C5 and the question papers have two parts (A & B), with instructions to attempt at least two questions from each part. Category C1, C9 and C10 are different from the other seven categories in a way that each question paper has ten parts separated by "OR" (i.e., 100% choice). Another important result based on Table-2 is that the categories C1, C2, C3, C8, C9 and C10 have no parts where as category C4, C5 and C7 question paper have two parts and the category C6 type of question paper has three parts (i.e., A, B and C).

After analyzing all the categories given in Table-2,3 it is observed that all question papers are different and have different percentages of choice available. In order to overcome this problem the proposed style of question papers is given in Table-4. This proposal requires a student to study all of the course/ chapters, otherwise he /she will loose a part or full questions of the paper.

The total contents should be divided into ten equal portions for two parts (questions) form each portion. This proposal will require the students to study entire of the syllabus, otherwise risk the loss of a proportional number of questions to be attempted.

The proposed model will require a teacher to teach the entire course. As a result the students will also study complete course. This will minimize the inflation of marks in the results and also upgrade our education standard.

Category	Subjects	Total Q's given	Comp. Q's	Parts	Attpt. Any	Choice
C1	English(C,E)	5	-	-	5	5OUT OF 5
C2	Botany-A,B Chemistry-B	8	-	-	4	4out8
C3	Zology-A Physics-A Chemistry-A Geography-A Pol.Scince Law-A Education	10	-	-	5	5out10
C4	Zoology-B Physics-B	10	-	I,II 5		At least 2 questions from each part
C5	Math's-a economics History	10	-	A,B	5	Atleast 2 questions from each part
C6	Math's-B	10	-	A,B,C	5	No more than tow form each part
C7	Computer-A,B	10	-	A,B	5	3 from part A 2 from part B
C8	Statistics-A,B	10	1	-	5	4 out of nine.
C9	Islamiat	4	-	-	4	Or among part "a" "b" of all questions
C10	Pashto	7	-	-	7	Or among part "a" "b" of all questions.

Table 3. Present Q .Oaper Style of BA/B.Sc.University of Peshawar

Source: Examination Section University of Peshawar.

Questions.No	Part A	Part B	Choice
1	Portion1	Portion2	OR
2	Portion1	Portion2	
3	Portion3	Portion4	OR
4	Portion3	Portion4	
5	Portion5	Portion6	OR
6	Portion5	Portion6	
7	Portion7	Portion8	OR
8	Portion7	Portion8	
9	Portion9	Portion10	OR
10	Portion9	Portion10	

Table 4. The Proposed Model Question Paper for 10 Chapter /Portions

3 Conclusions

1. The inflationary trend in marks obtained by the students has to be controlled in a systematic way.
2. The style of the question papers requires changes so that both the teachers and the students should get involved in teaching / learning of the entire syllabus rather than limiting only to a few chapters of the syllabus.
3. In order to achieve the above-mentioned target a new model of the question paper has been proposed which will help us in streamlining the academic activities in our education institutions, as follows

Study hours will be increased,
Attendance of the students and teachers will be improved,
Class teaching will be strengthened,
Discipline in college hours and hostels will be improved,
Student-teacher relations will be enhanced,
Cheating in the examination halls will be minimized,
Students non-productive activities will be reduced,
Selective teaching and studies for examinations will be avoided.

References

Ahmad, Siraj Uddin., 2000. *Quality Education* National Book Foundation Karachi, Pakistan.

Badshah, Sareer., 2000. *The Level of Subjectivity in the Evaluation of Examination Papers-I*, Sarhad Journal of Agr. Vol. 16(2).

Brophy, J.E., 1997. *Teacher Behavior and its Effect* Journal of Educational Psychology.

Iqbal, M., 1981. *Education in Pakistan* Lahore: Aziz Publishers.

National Education, Policy 1992-1998. Ministry of Education.

Raja, R.S., 1991. *Education for the twenty-first Century,* UNESCO Bangkok.

UNESCO, 1977. Regional office for Education in Asia, *The Training of Teacher in View of Changing Trends,* Bullties, Bangkok.

On The Rank of Variance Covariance Matrices

Bazargan-Lari, A.

P11121

Department of Statistics, Shiraz University, Iran.

Abstract. It is assumed that one has a set of n individuals that are to be randomly assigned to q different treatment groups. There are v variables and for variable α, there are m_α different categories. The expected value of $F_{i\alpha j}$ (frequency of i^{th} category and j^{th} group for variable α), covariance of two random variables $F_{i\alpha j}$ and $F_{g\beta t}$ are computed. The mean vector of $F_{i\alpha j}$ and the rank of variance covariance matrix is investigated.

Keywords. Categorical Data, Variance Covariance Matrix, Rank of Matrices, Direct Sum, Kronecher Product, Trac, Idempotent matrix.

1 Introduction

Assume that we have a set of n individuals that are to be randomly assigned to q different treatment groups. There are v variables and each variable is measured by nominal scale. For variable α there are $m_\alpha(\alpha = 1, 2, \ldots, v)$ different catergories. Let $R_1, R_2, \ldots, R_n$ denote the random variables with following joint probability.

$$P(R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n) = \begin{cases} \frac{1}{n!} \text{ for each permutation } \{r_1, \ldots, r_n\} \\ \quad \text{of} \{1, 2, \ldots, n\} \\ 0 \quad \text{otherwise} \end{cases}$$

And $A_1, A_2, \ldots, A_q$ denote sets (treatment groups) which are a partition of $\{1, 2, \ldots, n\}$. Then $R_p \in A_j$ means that individual p goes in treatment group j. Finaly let $\mathcal{B}_{1\alpha}, \mathcal{B}_{2\alpha}, \ldots, \mathcal{B}_{m_\alpha \alpha}$ be a partition of $\{1, 2, \ldots, n\}$ then for each $\alpha, p \in \mathcal{B}_{i\alpha}$ means that for variable α individual p is in the ith category.

Let $F_{i\alpha j}$ be the number of individuals in the ith category of variable α, and in group j. We define

$$B_{i\alpha j}(p, R_p) = \begin{cases} 1 \; p \in \mathcal{B}_{i\alpha} \text{ and} R_p \in A_j \\ 0 \text{ otherwise} \end{cases}$$

Then, for $i = 1, 2, \ldots, m_\alpha, \alpha = 1, 2, \ldots, v$ and $j = 1, 2, \ldots, q$ we have

$$F_{i\alpha j} = \sum_{p=1}^{n} B_{i\alpha j}(p, R_p) \tag{1.1}$$

In section 2, the mean vector of $F_{i\alpha j}$, (frequency of ith category and jth group for variable α) and the variance covariance matrix of $F_{i\alpha j}$ is investigated. In section 3, the rank of the variance covariance matrix of section 2 is discussed. Section 4 contains three examples.

2 Mean Vector and Vriance Covariance Matrix

Since $R_1, R_2, \ldots, R_n$ takes on the values of each permutation of $\{1, 2, \ldots, n\}$ with equal probability, so

$$P(R_p = r) = \frac{1}{n} \tag{2.1}$$

and

$$P(R_p = r, R_s = t) = \begin{cases} \frac{1}{n} & p = s \text{ and } r = t \\ \frac{1}{n(n-1)} & p \neq s \text{ and } r \neq t \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

hence

$$\bar{F}_{i\alpha j} = E(F_{i\alpha j}) = \sum_{r \in A_j} \sum_{p \in \mathcal{B}_{i\alpha}} P(R_p = r) = \sum_{r \in A_j} \sum_{p \in \mathcal{B}_{i\alpha}} \frac{1}{n}$$

If $N(\mathcal{B}_{i\alpha})$ denote the number of elements in the set of $\mathcal{B}_{i\alpha}$, and $N(Aj)$ denote the number of elements in the set of A_j, then

$$\bar{F} = \frac{1}{n} N(\mathcal{B}_{i\alpha}) N(A_j) \tag{2.3}$$

and

$$\begin{aligned} Cov(F_{i\alpha j}, F_{g\beta h}) = \tfrac{1}{n^2(n-1)} \left[nN(\mathcal{B}_{i\alpha} \cap \mathcal{B}_{g\beta}) - N(\mathcal{B}_{i\alpha}) N(\mathcal{B}_{g\beta}) \right] \\ \times \left[nN(A_j \cap A_h) - N(A_j) N(A_h) \right] \end{aligned} \tag{2.4}$$

The four subscripted set of values

$$N(\mathcal{B}_{i\alpha} \cap \mathcal{B}_{g\beta}) - \frac{1}{n} N(\mathcal{B}_{i\alpha}) N(\mathcal{B}_{g\beta})$$

can be arranged into a two dimensional array and denote it by matrix C, i.e.

$$C = \left[N(\mathcal{B}_{i\alpha} \cap \mathcal{B}_{g\beta}) - \frac{1}{n} N(\mathcal{B}_{i\alpha}) N(\mathcal{B}_{g\beta}) \right]$$

which its dimension is $\sum_{\alpha=1}^{v} m_\alpha$ by $\sum_{\alpha=1}^{v} m_\alpha$. Similarly we define matrix

$$T = \left[N(A_j \cap A_h) - \frac{1}{n} N(A_j) N(A_h) \right]_{q \times q}$$

Then the variance covariance matrix of $F_{i\alpha j}$ can be written as:

$$V = \frac{1}{n-1} C \otimes T \tag{2.5}$$

In which $\otimes$ denotes the kronecher product.

3 Rank of Variance Covariance Matrix

The purpose of this section is to find the rank of variance covariance matrix of $F_{i\alpha j}$. The following theorem will be useful.

Theorem 3.1: Let matrix M be written as follows:

$$M = \left[m_i^{\frac{1}{2}} m_j^{\frac{1}{2}} \delta_{ij} - \frac{1}{n} m_i m_j \right]_{k \times k} \tag{3.1}$$

in which $\sum_{i=1}^{k} m_i = n, m_i > 0$ and $\delta_{ij} = \begin{cases} 1 \; i = j \\ 0 \; i \neq j \end{cases}$, Then

$$rank(M) = k - 1 \tag{3.2}$$

Proof: Let D be a diagonal matrix with diagonal elements $m_i^{-\frac{1}{2}}$, i.e.

$$D = Dia \left(m_1^{-\frac{1}{2}} m_2^{-\frac{1}{2}} \dots m_k^{-\frac{1}{2}} \right)$$

Then we construct matrix $T = DMD$ which reduces to $T = I_k - P$, in which $P = \frac{1}{n} \underline{p} \, \underline{p}^{'}$ and $\underline{p}^{'} = \left[p_1^{\frac{1}{2}} p_2^{\frac{1}{2}} \dots p_k^{\frac{1}{2}} \right]$. It is easy to show that P is a symmetric idempotent matrix, so T is also symmetric idempotent matrix and hence

$$rank(T) = Trace(T) = Trace(I_k - P) = \sum_{i=1}^{k} (1 - \frac{m_i}{n}) = k - 1$$

and since D is a nonsingular matrix, so rank of M is equal to rank of T and the proof is completed.

Now we try to find the rank of matrix C. In order to do that, lets construct a quadratic form with matrix C and column vector X.

$$X^{'}CX = \sum_{\alpha=1}^{v}\sum_{\beta=1}^{v}\sum_{g=1}^{m_\beta}\sum_{i=1}^{m_\alpha}\left(N(\mathcal{B}_{i\alpha}\cap\mathcal{B}_{g\beta}) - \frac{1}{n}N(\mathcal{B}_{i\alpha})N(\mathcal{B}g\beta)\right)x_{g\beta}x_{i\alpha}$$

We put the follwoing distribution on the space consisting of the actual observation

$$\hat{f}_{\alpha\beta}(i,g) = \frac{1}{n}N(\mathcal{B}_{i\alpha}\cap\mathcal{B}_{g\beta})$$

which is bivariate probability density function and then the marginal probability density function is:

$$\hat{f}_{\alpha}(i) = \frac{1}{n}N(\mathcal{B}_{i\alpha})$$

so $X^{'}CX$ becomes

$$\begin{aligned} X^{'}CX = &\sum_{\alpha=1}^{v}\sum_{\beta=1}^{v}\sum_{g=1}^{m_\beta}\sum_{i=1}^{m_\alpha} N(\mathcal{B}_{i\alpha}\cap\mathcal{B}_{g\beta})\left[x_{i\alpha} - \tfrac{1}{n}\sum_{i=1}^{m_\alpha}N(\mathcal{B}_{i\alpha})x_{i\alpha}\right] \\ &\times\left[x_{g\beta} - \tfrac{1}{n}\sum_{g=1}^{m_\beta}N(\mathcal{B}_{g\beta})x_{g\beta}\right] \end{aligned} \tag{3.3}$$

Now consider

$$X^{'}CX = n\sum_{\alpha=1}^{v}\sum_{\beta=1}^{v}Cov(X_{i\alpha},X_{j\beta}) = n\sum_{\alpha=1}^{v}\sum_{\beta=1}^{v}\hat{\sigma}_{\alpha\beta}(x)$$

where x represent the assignment of scores to categories. Then $X^{'}CX = 0$ implies that there are certain relationships between the variables. Let

$$O_j^{'} = \left[O_{j_1}1, O_{j_1}2, \ldots, O_{j_v}v\right], \qquad j = 1,2,\ldots,n$$

denote the vectors of observed categories for each variable. If each $O_{j_i}i$ is replaced by the corresponding value $X_{j_i}i$, on obtains a transformed set of observations.

$$O_j^{*'} = \left[X_{j_1}1, X_{j_2}2, \ldots, X_{j_v}v\right], \qquad j = 1,2,\ldots,n$$

Then $X^{'}CX = 0$ implies that for each $j = 1,2,\ldots,n$, $\sum_{i=1}^{v}X_{j_i}i$ is constant. Since rank of C is equal to dimension of C minus dimension of

orthocomplement of C, and in most cases the maximum rank of C will be all that is necessary, so we will find the maximum rank of C, which is equal to dimension of C minus the minimum dimensin of orthocomplement of C. Thus, in order to find the minimum dimension of the orthocomplmente of C, we construct v linearly independent vectors of size $\sum_{\alpha=1}^{v} m_\alpha$ by 1 as follows:

$$\begin{gathered}
[X_{11}, X_{21}, \ldots, X_{m_1 1}, 0, 0, \ldots 0, 0, 0, \ldots, 0]^{'} \\
[0, 0, \ldots, 0, X_{12}, X_{22}, \ldots, X_{m_2 2}, 0, 0, \ldots 0,]^{'} \\
\cdots \qquad \cdots \qquad \cdots \\
[0, 0, \ldots, 0, 0, 0, \ldots, 0, X_{1v}, X_{2v}, \ldots, X_{m_v v}]^{'}
\end{gathered}$$

where $X_{ji} = X_{gi} = k_i$ and $j, g = 1, \ldots m_i$. These vectors are linearly independent because they are pairwise orthogonal. The above v vectors are in the orthocomplement of C, that is each vector is orthogonal to C, because it is possible to show that $CX = 0$. Therefore, in the orthocomplement of C there are v linearly independent vectors which implies that the dimension of orthocomplement can not be less than v, or a lower bound for dimension of orthocomplement of C is equal to v. Finally

$$\text{rank}(C) \leq \sum_{\alpha=1}^{v} m_\alpha - v = \sum_{\alpha=1}^{v} (m_\alpha - 1) \tag{3.4}$$

Now we show that $\sum_{\alpha=1}^{v} (m_\alpha - 1)$ is the maximum rank of the matrix C, to prove that, it is sufficient to show that there exists a matrix C with the exact rank equal to $\sum_{\alpha=1}^{v} (m_\alpha - 1)$. To construct such C, we make the following assumptions:

1: If $\alpha \neq \beta$ then $N(\beta_{i\alpha} \cap \mathcal{B}_{g\beta}) = \frac{1}{n} N(\mathcal{B}_{i\alpha}) N(\mathcal{B}_{g\beta})$

2: If $\alpha = \beta$ then $N(\beta_{i\alpha} \cap \mathcal{B}_{g\beta}) = \delta_{ij} N(\mathcal{B}_{i\alpha})$

3: For all i and α, $N(\mathcal{B}_{i\alpha}) \neq 0$

with these assumption the matrix C becomes a block diagonal matrix where each matrix on the diagonal has the property that the sum of each row and each column is zero. In matrix notation, we can write C as follows:

$$C = Diag[P_\alpha], \qquad \alpha = 1, 2, \ldots, v$$

where P_1 is m_1 by m_1 and P_2 is m_2 by m_2 and so on. Now each P_α's, $\alpha = 1, 2, \ldots, v$ can be written as:

$$P_\alpha = \left[\delta_{ig} N^{\frac{1}{2}}(\mathcal{B}_{i\alpha}) N^{\frac{1}{2}}(\mathcal{B}_{g\beta}) - \frac{1}{n} N(\mathcal{B}_{i\alpha}) N(\mathcal{B}_{g\beta})\right]$$

where dimension of P_α is m_α by m_α and $\sum_{i=1}^{m_\alpha} N(\mathcal{B}_{i\alpha}) = n$.

The rank of matrix C is equal to the sum of the ranks of the P_α's. So by using the theorem (3.1) the rank of each P_α's, $\alpha = 1, \ldots, v$ is equal to $(m_\alpha - 1)$ for $\alpha = 1, 2, \ldots, v$. Hence

$$\text{rank}(C) = \sum_{\alpha=1}^{v} \text{rank}(P_\alpha) = \sum_{\alpha=1}^{v} (m_\alpha - 1)$$

It is easy to show that the rank of matrix T is equal to $(q - 1)$. So from (2.5) we conclude that the rank of V is equal to rank of T multiply by rank of C. So the maximum rank of V is as follows:

$$\max . rank(V) = (q-1) \left[\sum_{\alpha=1}^{v} (m_\alpha - 1)\right] \tag{3.5}$$

Remark 3.1: In the most case the exact rank of variance covariance matrix is equal to the maximum rank, except for the case that there exists a certain linear relationship between variables, or in the case in which the sample size is small. Weeks and Williams (1964), have given a sufficient condition for having the rank of V attain its maximum.

4 Examples

Example 4.1: A survey was being conducted on the prevalence of pneumoconioses among miners. The miners has chest X-ray taken. These X-ray films were then submitted to three doctors. The different machines A, B and C were used to read the X-ray films. These machines remained stationary. Each of the machines were situated in a different location far from each other. The doctors with no consultation among themselves, categorized the films according to the three categories:

0- No disease

1- Stage 1 and 2 simple pneumoconiosis

2- Stage 3 simple pneumoconiosis

A sample of thirty X-ray films were taken, and were randomly assigned to three groups 1, 2 and 3. Each groups consisted of ten films. The films comprising group 1 were sent to the doctors and each of the three doctors read the group 1 films on machine A. Likewise, groups 2 films were read on machine B by all three doctors and group 3 films were read on machine C by all three doctors. The assignment of these films to three groups is given in the following table:

	Doctors				Doctors				Doctors			
	Film	M	P	Q	Film	M	P	Q	Film	M	P	Q
	7	1	1	1	12	1	2	1	29	1	2	1
	5	1	1	0	8	0	0	0	3	1	0	0
	20	1	2	1	24	1	2	2	11	1	2	1
Group 1	22	2	2	2	Group 2 2	1	0	1	Group 3 6	1	0	0
	30	1	1	0	9	0	0	0	10	1	2	2
	28	1	2	0	16	0	0	0	21	1	1	0
	19	1	2	1	13	0	1	0	25	0	1	1
	4	1	0	0	17	1	2	1	27	2	2	2
	1	1	1	0	28	0	1	0	14	1	2	1

then

$$F_{i\alpha j} = \begin{bmatrix} 0 \\ 9 \\ 1 \\ 1 \\ 5 \\ 4 \\ 5 \\ 4 \\ 1 \\ 5 \\ 4 \\ 1 \\ 4 \\ 2 \\ 4 \\ 6 \\ 3 \\ 1 \\ 1 \\ 8 \\ 1 \\ 2 \\ 3 \\ 5 \\ 3 \\ 5 \\ 2 \end{bmatrix} \qquad \bar{F}_{i\alpha j} = \begin{bmatrix} 2.0 \\ 7.0 \\ 1.0 \\ 2.3 \\ 3.3 \\ 4.3 \\ 4.7 \\ 4.0 \\ 1.3 \\ 2.0 \\ 7.0 \\ 1.0 \\ 2.3 \\ 3.3 \\ 4.3 \\ 4.7 \\ 4.0 \\ 1.3 \\ 2.0 \\ 7.0 \\ 1.0 \\ 2.3 \\ 3.3 \\ 4.3 \\ 4.7 \\ 4.0 \\ 1.3 \end{bmatrix}$$

and the matrices C and T are computed as follows:

$$C = \begin{bmatrix} 4.8 & -4.2 & -0.6 & 1.6 & 1.0 & -2.6 & 2.2 & -1.4 & -0.8 \\ -4.2 & 6.3 & -2.1 & -0.9 & 0.0 & 0.9 & -1.8 & 2.6 & -0.8 \\ -0.6 & -2.1 & 2.7 & -0.7 & -1.0 & 1.7 & -0.4 & -1.2 & 1.6 \\ 1.6 & -0.9 & -0.7 & 5.3 & -2.3 & -3.0 & 2.7 & -1.8 & -0.9 \\ 1.0 & 0.0 & -1.0 & -2.3 & 6.6 & -4.3 & 1.3 & 0.0 & -1.3 \\ -2.6 & 0.9 & 1.7 & -3.0 & -4.3 & 7.3 & -4.0 & 1.8 & 2.2 \\ 2.2 & -1.8 & -0.4 & 2.7 & 1.3 & -4.0 & 7.4 & -5.6 & -1.8 \\ -1.4 & 2.6 & -1.2 & -1.8 & 0.0 & 1.8 & -5.6 & 7.2 & -1.6 \\ -0.8 & -0.8 & 1.6 & -0.9 & -1.3 & 2.2 & -1.8 & -1.6 & 3.4 \end{bmatrix}$$

$$T = \begin{bmatrix} 6.6 & -3.3 & -3.3 \\ -3.3 & 6.6 & -3.3 \\ -3.3 & -3.3 & 6.6 \end{bmatrix}$$

Then $V = \frac{1}{29} T \otimes C$ and its rank is equal to $(3-1)(\sum_{i=1}^{3}(3-1)) = 2 \times 6 = 12$.

Example 4.2: In theorem 3.1, if $m_1 = m_2 = \ldots = m_k = 1$, then $M = I_k - \frac{1}{k} J_k$ in which J_k is a square matrix that all its elements are equal to one. In this case matrix M is called centring matrix and is denoted by C. Then rank of C is equal to $(k-1)$ which is degrees of freedom for distribution of quadratic form $X^{'}CX$. Also it is useful for rank of wishart matrix.

Example 4.3: Let M_i for $i = 1, 2, \ldots N$ be a square matrix k_i by k_i which satisfy conditions of theorem 3.1, and if

$$D = M_1 \oplus M_2 \oplus \ldots \oplus M_N = Diag(M_1 M_2 \ldots M_N)$$

which $\oplus$ denotes direct sum, then

$$rank(D) = \sum_{i=1}^{N} rank(M_i) = \sum_{i=1}^{N} (k_i - 1)$$

To illustrate example 4.3 consider the following matrix

$$D = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{bmatrix}$$

then $rank(D) = (2-1) + (3-1) + (4-1) = 6$

5 Acknowledgment

The author wishes to thank the research council of shiraz university for their financial support.

References

Graybill. F. A. (1983). *Matrices with Applications in Statistics.* Wadsworth, california.

Hohn, F. (1958). *Elementary Matrix Algebra.* Macmillian, New York.

Hogg, R. V. and Craig, A. T. (1978). *Introduction to Mathematical Statistics.* Macmillian, New York.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics,* John wiley & Sons, New York.

Snedecor, G. W. and Cochran, W. G. (1973). *Statistical Methods,* The Iowa State University Press, Ames.

Weeks, L. D. and Williams, R. D. (1964) *On Connectedness, Technometrics,* 6, 319-324.

Comparison of homogeneous and heterogeneous modelling in Longitudinal data via the use of smoothing splines

Borhani, H. A.[1] **and Davies, R. B.**[2]

P11002

[1] Department of Statistics, Shiraz University, Iran.
[2] Lancaster University, UK.

Abstract. The theory of smoothing splines in the context of semi-parametric generalized linear models is well developed(e.g. Green and Silverman, 1994), but applications tend to raised computational and pragmatic problems. In this paper an example is considered and we compare the use of smoothing splines in logistic regression model, having discretised event history sequences both with and without random effect specification to control for omitted variables. The main emphasis is upon using cubic spline functions for a number of temporal variables.

Keywords: Exponential Family Distribution, Spline Smoothing Function, Penalized Likelihood, Homogeneity, Heterogeneity, Akaike's Information Criterion.

1 Introduction

It is often important to identify the functional relationship between a number of temporal variables and the response variable in modelling panel and other longitudinal micro-level data. It can also be important to avoid spurious time dependencies by controlling for the effect of variables omitted from the analyses. In this paper we are interested in disentangling the simultaneous effects of three temporal variables: age, year, and duration of stay and with using cubic smoothing splines to characterise the individual effects. We investigate a non-parametric approach because temporal effects often vary in complex ways. Modelling temporal variables needs to allow for residual heterogeneity; it is well known that results for duration and other temporal effects can be seriously biased if residual heterogeneity is ignored (Lancaster and Nickel, 1980 and Heckman and Singer, 1985).

Migration is a relevant process to study because there is evidence that all the temporal effects mentioned above are relevant. Migration behaviour is characterised by strong temporal dependencies. A decision to migrate depends upon the interval since a previous move. In particular, migration is a duration dependent with evidence of inertial effects(Dale and Davies, 1994). A comparison by the use of cubic smoothing splines in logistic regression models is performed both with and without random effect specifications to control for omitted variables when the migration history sequences are discretised. We adopt an empirical, investigative approach using a dataset on inter-county migration, we also examine the further complications which arise in fitting univariate splines.

2 The Data

The migration history data were extracted from a large retrospective survey of life and work histories conducted in the U.K. in 1986. The analyses presented are from male respondents from a specific area in the north of England. Migration is defined as a move between counties. The sequence for each individual commences with their first job.

3 Methodological Approach

Let Y_{ij} denote the outcome or the binary response of the i-th individual at time j, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, T_i$. Y_{ij} takes value 1 if individual i has a move at time j and 0 otherwise.

The explanatory variables, X_{ij}, consist of age (A), year-1900 (Y), and duration of residence (D) of respondents at each time interval. Interest focuses on the functional relationships between these variables and Y_{ij}. Initially we model two of the variables parametrically and, in turn, the third variable non-parametrically using a smoothing spline function. Let the probability of move for the individual i at time j be p_{ij}, with $Y_{ij} \sim B(1, p_{ij})$. Also assume the monotone differentiable link function G is a logit link such that $G(p_{ij}) = x_{ij}^T\beta + g(t_{ij})$, where $p_{ij} = E(Y_{ij}|X_{ij})$, β is a p-vector of unknown parameters and g is a smooth real function of the splined variable, t. If all t_{ij} are not distinct, we can construct a matrix N to transform t_{ij} into a set of distinct

values s_k, where $k = 1, 2, \ldots, q$. So instead of t_{ij} we put s_k and in the linear predictor g is replace by Ng. Therefore the likelihood for both homogeneous and heterogeneous models are:

3.1 Homogeneous Model

The likelihood for the ith individual becomes

$$L_i = \prod \lim_{j=1}^{T_i} \frac{\left\{exp\left[x_{ij}^T\beta + (Ng)_{(j)}\right]\right\}^{y_{ij}}}{1 + exp\left[x_{ij}^T\beta + (Ng)_{(j)}\right]}.$$

The log likelihood can be written as, $l(\beta, g) = \sum_{i=1}^{n} \log(L_i)$. Following Good and Gaskin (1971), Silverman (1985), and Green (1987) we use the penalized likelihood that should be maximized over all β and g and is given by

$$l_p(\beta, g) = l(\beta, g) - \frac{1}{2}\lambda \int_a^b g''^2(t)d(t), \tag{1}$$

where $\int_a^b g''^2$ denotes a measure of rapid local variation and the smoothing parameter λ controls smoothness of the estimated g, $0 < \lambda < \infty$. As λ tends to zero the graph of $\widehat{g}$ shows many rapid fluctuations. Whereas, as λ tends to infinity the estimate of g tends to a linear curve.

3.2 Heterogeneous Model

We allow for unobserved heterogeneity in migration behaviour by including an unknown nuisance parameter, e_i, in the linear predictor of the model. Inference is based upon the integrated likelihood

$$l_m(\beta, g) = \sum_{i=1}^{n} log\left\{\int \prod_{j=1}^{T_i} \frac{\left\{exp\left[x_{ij}^T\beta + (Ng)_{(j)} + e_i\right]\right\}^{y_{ij}}}{1 + exp\left[x_{ij}^T\beta + (Ng)_{(j)} + e_i\right]} f(e_i)\, de_i\right\}, \tag{2}$$

Variable	Model 1 : Homogenous		Model 2 : Heterogenous	
	P. E.	S.E.	P. E.	S.E.
Duration	-1.070	0.080	-0.073	0.100
Year	-0.041	0.006	-0.051	0.009
Scale	-	-	0.920	0.132
λ	19.5		19.5	
Deviance	2285.90		2259.90	

Table 1. Model Fitting Results with Smoothing Spline for Age

where f(e_i) is the density function of e_i. Assuming that $e_i \sim N(0, \sigma^2)$, the integral in (2) can be calculated numerically using standard quadrature methods. The corresponding penalized log likelihood is given by $l_{mp}(\beta, g) = l_m(\beta, g) - \frac{1}{2}\lambda \int_a^b g''^2(t)d(t)$.

4 Model Fitting

Following Marx and Eiler (1996) we investigate the use of Akaike's Information Criterion, AIC, to derive an empirically reasonable value for the smoothing parameter λ. Akaike's Information Criterion for a fixed parameter λ is defined as:

$$AIC(\lambda) = Deviance(y, \hat{\beta}, \hat{g}, \lambda) + 2dim(\hat{\beta}, \hat{g}, \lambda), \tag{3}$$

where $dim(\hat{\beta}, \hat{g}, \lambda) = Trace(A)$, $A = S + S_1$, $S = N(N^T W N + \lambda K)^{-1} N^T W$,

$S_1 = (I-S)X\left\{X^T W(I-S)X\right\}^{-1} X^T W(I-S)$, S is the smoother or hat matrix, K is a fixed and W is a diagonal matrix. For model selection, after maximizing the penalized log likelihood over a range of different values for λ we select the λ which minimizes the AIC.

Three temporal explanatory variables are under consideration: Two in the parametric part of the model and the third in the nonparametric part. Also, two types of the models are of interest : homogeneous and heterogeneous.

The results using a spline representation for Age are given in Table 1. Duration of stay and year are included as linear terms. Moreover, computation of the AIC was simplified by specifying $A = S$ in (3) .

As expected, the random effects specification gives a substantial improvement in the model deviance. Also, controlling for omitted variables in this way attenuates substantially the cumulative inertia (negative duration-of-stay) effects. The spline results for $\lambda = 1, 19.5, 30$ are also illustrated graphically in Figure 1 with year set at 1986 and duration of stay at 10 years. As expected, the graph of probability against age becomes smoother in both models as λ increases. These figures reveal three peaks. This is a novel finding, perhaps because no one has attempted to fit polynomials of order 6 to migration data, and raises interesting substantive questions.

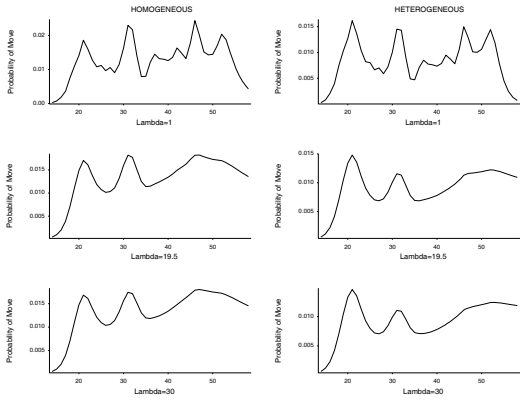

Fig. 1. Spline representation for age

5 Advantages and disadvantages of AIC

Although AIC attains its minimum value at different λ's in homogeneous and heterogeneous models the values tend to be similar and it appears that the λ which minimises AIC in a homogeneous model is an adequate approximation for the corresponding heterogeneous model (See Tables 2 and 3). Also, visual inspection of plots (as in Figure 1) confirm that the simplified AIC gives plausible results for age and

Homogeneous Model		
λ	**Simplified AIC**	**Full argumented AIC**
1	2315.72138	2319.70893
10	2307.46256	2311.45621
19.47	2306.86344	2310.85577
20	2306.86510	2310.85721
30	2307.17283	2311.16392

Table 2. Comparison of the values of AIC

Heterogeneous Model		
λ	**Simplified AIC**	**Full argumented AIC**
10.0	2281.18966	2285.18436
19.47	2280.66969	2284.66446
70.0	2283.69634	2287.68567

Table 3. Comparison of the values of AIC

duration of stay. The full AIC (i.e. with $A = S + S_1$) requires substantially more computation but was found to attain its minimum at the same λ. This is because Trace(S_1) is relatively insensitive to λ and is approximately equal to the number of parameters in the parametric part of the linear predictor, for both homogenous and heterogenous models.

However, the AIC was not successful in identifying an appropriate value of the smoothing parameter λ when year was used as the splined variable. This contrasted with the visual impression that at approximately $\lambda = 100$ there was a compromise smoothness with irregular fluctuations smoothed out but more systematic variations still evident. Other authors have noted that statistical criteria for choosing a smoothing parameter are not always as effective as visual methods (e.g. Hastie and Tibshirani, 1990).

6 Other Issues

The AIC was found to give identical optimum λ values for corresponding homogenous and heterogenous specifications. The same result was noted by Chesher (1997). As the heterogenous models are so computationally excessive, this is an important result. It enables the repeated model fitting to locate the minimum AIC to be confined to the

computationally simpler homogenous model with just one fit of the heterogenous model when the optimum λ has been identified.

To investigate the sensitivity of the smoothing splines for one variable to the parametric representation of the others, we fitted polynomials of degree 4 for duration of stay and year, using age as a spline variable. The AIC again gave $\lambda = 19.5$ and there was encouragingly little change in the plot of probability of move against age (see Figure 2). In particular, the three peaks remained; they do not appear to be an artifact of misspecification of the other two variables.

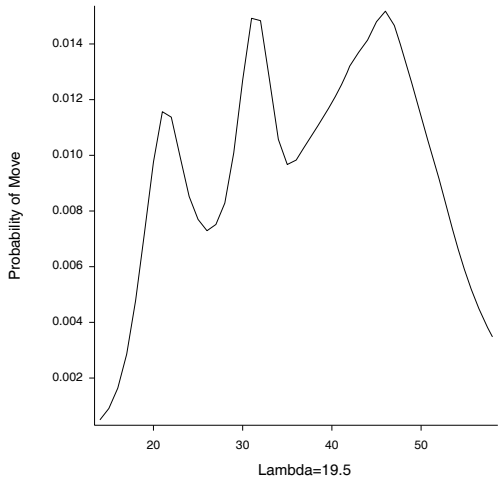

Fig. 2. Comparison of values of AIC

We fitted the same polynomials of the same degree for duration-of-stay and year with age as a spline variable with the optimum $\lambda = 19.5$ in a heterogeneous model corresponding to the homogeneous model (see Figure 3). Random effect specification does not alter the shape.

7 Concluding comments

The use of smoothing splines is computationally more demanding for heterogenous (random effect) models than for the corresponding homogenous specification but they do not appear to pose any additional

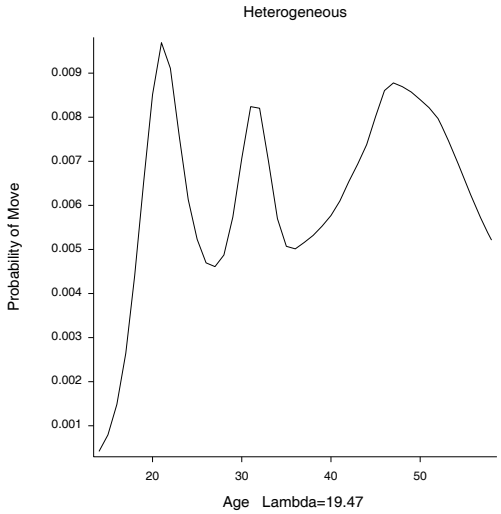

Fig. 3. Smoothing spline representation for age in heterogeneous model with 4th degree polynomial for the other variables

problems. Moreover, computational effort can be reduced by identifying the appropriate value of the smoothing parameter from repeated fitting of the homogenous model.

The effectiveness of smoothing spline methods is demonstrated by the application described in the paper. Not only do we identify an unexpectedly complex relationship between age and residential mobility but we also show that this result is unlikely to be due to misspecification of the parametric part of the model.

8 Acknowledgments

The first author thanks the research council of Shiraz University.

References

Chesher, A. (1997). Diet Revealed ?: Semiparametric Estimation of Nutrient Intake-Age Relationships. *J.R.Statist.Soc. A(1997)* Forthcoming.

Dale, A. and Davies, R.B. (1994). Analyzing social and politi-

cal change. SAGE publications Ltd, London.

Good, I. J. and Gaskins, R.A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58**, 255-277.

Green, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *Int. Statist. Rev.* **55**, 245-60.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric regression and generalized linear models*. Chapman and Hall, London.

Hastie,T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Heckhman, J. J. and Singer, B. (1985). *Longitudinal Analysis of Labor Market Data, p. 39- 58.* Cambridge University Press, Cambridge.

Lancaster, T. and Nickle, S. (1980). The analysis of re-employment probabilities for the unemployed. *J. R. Statist. Soc. A. Part, 2, p. 141-165.*

Marx, B.D. and Eiler, P.H. (1996). Generalized linear regression on sample signal with penalized likelihood. *Proceedings of the 11th International workshop on statistical modelling.* 259-266**.**

Silverman, B.W. (1985). Penalized maximum likelihood estimation. In *Encyclopedia of Statistical Sciences,* ***6****,* Ed. S. Kotz and N.L. Johnson, pp. 664-667. Wiley, New York.

Empirical Bayes Analysis of Generalized Logistic Regression Models for Multinomial Responses

Farzad Eskandari, Mohammad R.Meshkani

P11142

Department of Statistics, Shahid Beheshti University, Iran.

Abstract. In this paper, we develop an empirical Bayes approach for estimation of a generalized logistic regression model with repeated expriments. For a hypothesis concerning the parameters of a logistic regression model, first we compute the exact posterior distribution coresponding to the conjugate prior distribution where in the superparametrs have been estimated by the method of moments and Maximum likelihood method. Finally, to describe the relationship between responce vector and covariates, we estimate vector of $\boldsymbol{\beta}$, via itterative empirical Bayes approach. Following Bayesian paradigm, the Bayes and empirical Bayes estimators relative to various loss functions are obtained. These procedures are illustrated by a real example.

Keywords. Logistic models, Multinomial distribution, Bayes, Empirical Bayes, Model Selection

1 Introduction

In many applications, the response of each subject is measured at several occasions, for instance at several time points or under several conditions belonging to one and only one of certain distinct categories. As an example, consider the situation where we wish to study tumor type (embryonal, alveolar, pleomorphic) for patients with rhabdomyosar coma. Furthermore, there may be some covariates of interest such as, age($''0'' \equiv\leq$ 15 years, $''1'' \equiv>$ 15 years) and sex($''0'' \equiv$ male, $''1'' \equiv$ female). Usually the response is observed for each subject at I occasions or I locations, and interest centers on the relationship between the response variable and the covariates. As another real-life example, Schmidt and Strauss (1975) modeled the occuptional attainment in the United States, using covariates such as years of schooling, labor market

experience (calculated as age-years of schooling -5), race(1≡ white, 0≡ black), and sex(1≡ male, 0≡ female). The categories of occupational attainment are professional, white collar, blue collar, craft, and menial. In another application, Schull(1958) studied pregnancy outcome in three districts of Shizuoka city, Japan, according to the degree of consanguinity between the parents. In his study, death (categorized as abortion, stillbirth, in less than 12 months, in 13-60 months, survived) is considered as a multinomial response variable, and residence(Rural district, Intermediate district, Urban district) and Consanguinity (no relation, 2nd cousins, 1st cousins) are covariates. Alternatively, Forster (1999) developed Metropollis-Hastings algorithm for exact inference on binomial and multinomial logistic regression models based on repeated categorical response. In this paper, these types of problems are studied via the empirical Bayes approach.
Explicity, we consider the logistic regression analysis for a multinomial response variable. In Section 2, the model is presented. Section 3 provides the posterior distribution of the regression parameters. In Section 4, Bayes and empirical Bayes estimates are obtained. Section 5 contains the model selection procedure. Finally, in Section 6, a real data set is analysed.

2 The Model

Consider a multinomial response variable $\mathbf{Y}$ with categories $0, 1, \cdots, K$ where for $i = 1, 2, \cdots, I$, the vector $\mathbf{Y}_i \sim Mult(N_i, \mathbf{P}_i)$. The ith count vector is denoted by $(y_{i0}, \cdots, y_{iK})$, whose total count $N_i = \sum_{j=0}^{K} y_{ij}$ is assumed to be fixed. Let

$$\mathbf{Y} = \begin{pmatrix} y_{10}\ y_{11} \cdots y_{1K} \\ y_{20}\ y_{21} \cdots y_{2K} \\ \vdots\ \ \vdots\ \ \ldots\ \ \vdots \\ y_{I0}\ y_{I1} \cdots y_{IK} \end{pmatrix} = \Big(\mathbf{y}_0, \mathbf{y}_1, \cdots, \mathbf{y}_K\Big),$$

denote an $I \times (K+1)$ matrix of responses whose columns are $\mathbf{y}_1$, $\mathbf{y}_2$, $\cdots$, $\mathbf{y}_K$ with $\mathbf{y}_0 = \mathbf{N} - \sum_{j=1}^{K} \mathbf{y}_j$, and $\mathbf{N} = (N_1, \cdots, N_I)'$. The likelihood fuction is

$$L(\mathbf{P}) = \Big(\prod_i \frac{N_i!}{\prod_j y_{ij}!}\Big) exp\Big(\sum_i \sum_j y_{ij} Ln(P_{ij})\Big),$$

i.e., $\mathbf{Y} \sim Product\ Mult(\mathbf{N}, \mathbf{P})$, with

$$\mathbf{P} = \begin{pmatrix} p_{10}\ p_{11} \cdots p_{1K} \\ p_{20}\ p_{21} \cdots p_{2K} \\ \vdots \quad \vdots \quad \cdots \quad \vdots \\ p_{I0}\ p_{I1} \cdots p_{IK} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \\ \vdots \\ \mathbf{P}_K \end{pmatrix},$$

and $\sum_{j=0}^{K} p_{ij} = 1$. Suppose, for $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, K$, we set

$$H : \eta_{ij} = Ln\Big(\frac{N_i P_{ij}}{N_i P_{i0}}\Big). \qquad (2-1)$$

Then, the canonical form of the likelihood function with respect to (2-1) will be

$$L(\boldsymbol{\eta}) = \Big(\prod_i \frac{N_i!}{\prod_j y_{ij}!}\Big) exp\Big\{\sum_i \sum_j y_{ij}\eta_{ij} - \sum_i N_i Ln\Big(\sum_j exp(\eta_{ij})\Big)\Big\}. \qquad (2-2)$$

The conjugate prior distribution for the vector of $\mathbf{P}_i$ is a Dirichlet distribution with superparameter $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_K)$. To determine the prior distribution for the canonical form (2-2), one should modify the Dirichlet distribution according to the transformation (2-1). From (2-1), the joint distribution of $\boldsymbol{\eta_i}$ is obtained as

$$\pi(\boldsymbol{\eta}) = \Big(\prod_i \frac{\Gamma(\alpha_.)}{\prod_j \Gamma(\alpha_j)}\Big) exp\Big\{\sum_i \sum_j \alpha_j \eta_{ij} - \alpha_. \sum_i Ln\Big(\sum_j exp(\eta_{ij})\Big)\Big\}. \qquad (2-3)$$

with $\alpha_. = \sum_j \alpha_j$.

3 The Posterior Distribution

First, we compute the marginal distribution of $\mathbf{Y}$, assuming $\boldsymbol{\alpha}$ is known. From (2-2) and (2-3), we have

$$f(\mathbf{y}|\boldsymbol{\alpha}) = \int_{\boldsymbol{\eta}} L(\boldsymbol{\eta}) \times \pi(\boldsymbol{\eta}) d\boldsymbol{\eta}$$

$$= \int_{\boldsymbol{\eta}} \prod_i \frac{N_i!}{\prod_j y_{ij}!} exp\Big\{\sum_i \sum_j y_{ij}\eta_{ij} - \sum_i N_i Ln\Big(\sum_j exp(\eta_{ij})\Big)\Big\} \times$$

$$\prod_i \Big(\frac{\Gamma(\alpha_.)}{\prod_j \Gamma(\alpha_j)}\Big) exp\Big\{\sum_i \sum_j \alpha_j \eta_{ij} - \sum_i (\alpha_.) Ln\Big(\sum_j exp(\eta_{ij})\Big)\Big\} d\boldsymbol{\eta}.$$

$$= \prod_i \frac{N_i!}{\prod_j y_{ij}!}\Big(\frac{\Gamma(\alpha_.)}{\prod_j \Gamma(\alpha_j)}\Big)$$

$$\int_{\boldsymbol{\eta}} exp\Big\{\sum_i \sum_j (\alpha_j + y_{ij})\eta_{ij} - \sum_i (N_i + \alpha_.) Ln\Big(\sum_j exp(\eta_{ij})\Big)\Big\} d\boldsymbol{\eta}$$

$$= \prod_i \frac{N_i! \Gamma(\alpha_.) \prod_j \Gamma(\alpha_j + y_{ij})}{\prod_j y_{ij}! \prod_j \Gamma(\alpha_j) \Gamma(\alpha_. + N_i)}\Big)$$

Hence, the posterior distribution is equal to

$$\pi(\boldsymbol{\eta}|\mathbf{Y} = \mathbf{y}) = \frac{L(\boldsymbol{\eta}) \times \pi(\boldsymbol{\eta})}{\int_{\boldsymbol{\eta}} L(\boldsymbol{\eta}) \times \pi(\boldsymbol{\eta}) d\boldsymbol{\eta}}$$

$$= C.exp\Big\{\sum_i \sum_j (\alpha_j + y_{ij})\eta_{ij} - \sum_i (\alpha_. + N_i) Ln\Big(\sum_j exp(\eta_{ij})\Big)\Big\}. \quad (3-1)$$

where

$$C = \prod_i \Big(\frac{\Gamma(\alpha_. + N_i)}{\prod_j \Gamma(\alpha_j + y_{ij})}\Big).$$

Later, we need to compute the posterior moments of $\boldsymbol{\eta}$. However, the direct calculation of the posterior moments of $\boldsymbol{\eta}$ is somewhat intractable. Thus, for this purpose we shall use the posterior distribution of $\mathbf{P}$ which is known to have a Dirichlet distribution for each component. Hence, we have

$$E(\eta_{ij}|\mathbf{y}_i, \boldsymbol{\alpha}) = E\Big(Ln(P_{ij}|\mathbf{y}_i, \boldsymbol{\alpha})\Big) - E\Big(Ln(P_{i0}|\mathbf{y}_i, \boldsymbol{\alpha})\Big)$$

$$= \int Ln(p_{ij}) \frac{\Gamma(\alpha_. + N_i)}{\Gamma(\alpha_j + y_{ij})\Gamma(\alpha_. + N_i - \alpha_j - y_{ij})} p_{ij}^{\alpha_j + y_{ij} - 1} (1 - p_{ij})^{\alpha_. + N_i - \alpha_j - y_{ij} - 1} dp_{ij}$$

$$- \int Ln(p_{i0}) \frac{\Gamma(\alpha_. + N_i)}{\Gamma(\alpha_0 + y_{i0})\Gamma(\alpha_. + N_i - \alpha_0 - y_{i0})} p_{i0}^{\alpha_0 + y_{i0} - 1} (1 - p_{i0})^{\alpha_. + N_i - \alpha_0 - y_{i0} - 1} dp_{i0}$$

$$= \frac{\Gamma(\alpha_. + N_i)}{\Gamma(\alpha_j + y_{ij})\Gamma(\alpha_. + N_i - \alpha_j - y_{ij})} \int_{p_{ij}} \frac{\partial}{\partial \alpha_j} p_{ij}^{\alpha_j + y_{ij} - 1} (1 - p_{ij})^{\alpha_. + N_i - \alpha_j - y_{ij} - 1} dp_{ij}$$

$$- \frac{\Gamma(\alpha_. + N_i)}{\Gamma(\alpha_0 + y_{i0})\Gamma(\alpha_. + N_i - \alpha_0 - y_{i0})} \int_{p_{i0}} \frac{\partial}{\partial \alpha_0} p_{i0}^{\alpha_0 + y_{i0} - 1} (1 - p_{i0})^{\alpha_. + N_i - \alpha_0 - y_{i0} - 1} dp_{i0}$$

$$= \frac{\Gamma(\alpha_. + N_i)}{\Gamma(\alpha_j + y_{ij})\Gamma(\alpha_. + N_i - \alpha_j - y_{ij})} \frac{\partial}{\partial \alpha_j} \Big(\frac{\Gamma(\alpha_j + y_{ij})\Gamma(\alpha_. + N_i - \alpha_j - y_{ij})}{\Gamma(\alpha_. + N_i)}$$

$$-\frac{\Gamma(\alpha_. + N_i)}{\Gamma(\alpha_0 + y_{i0})\Gamma(\alpha_. + N_i - \alpha_0 - y_{i0})}\frac{\partial}{\partial\alpha_0}\frac{\Gamma(\alpha_0 + y_{i0})\Gamma(\alpha_. + N_i - \alpha_0 - y_{i0})}{\Gamma(\alpha_. + N_i)}).$$

Hence,

$$E(\eta_{ij}|\mathbf{y}_i,\boldsymbol{\alpha}) = \Psi(\alpha_j + y_{ij}) - \Psi(\alpha_0 + y_{i0}) - \frac{d}{d\alpha_j}\Psi\left(\alpha_. + N_i\right) + \frac{d}{d\alpha_0}\Psi\left(\alpha_. + N_i\right), \tag{3-2}$$

where $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$.
Similarly, the Covariance matrix of $(\boldsymbol{\eta}_i|\mathbf{y}_i)$ as function of $\boldsymbol{\alpha}$ is computable as

$$Cov\left(\boldsymbol{\eta}_i|\boldsymbol{\alpha},\mathbf{y}_i\right) = \Sigma_i(\boldsymbol{\alpha},\mathbf{y}_i) = \begin{pmatrix} \sigma_{1i}^2 & \sigma_{12i} & \cdots & \sigma_{1Ki} \\ \sigma_{21i} & \sigma_{2i}^2 & \cdots & \sigma_{2Ki} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{K1i} & \sigma_{K2i} & \cdots & \sigma_{Ki}^2 \end{pmatrix} \tag{3-3}$$

where

$$\sigma_{ji}^2 = Var(\eta_{ij}|\mathbf{y}_i) = \frac{d}{d\alpha_j}\Psi(\alpha_j + y_{ij}) + \frac{d}{d\alpha_0}\Psi(\alpha_0 + y_{i0})$$

and

$$\sigma_{jj'i} = Cov(\eta_{ij},\eta_{ij'}|\mathbf{y}_i) = \frac{d^2}{d\alpha_j d\alpha_0}\Psi\left(\alpha_. + N_i\right) - \frac{d^2}{d\alpha_{j'} d\alpha_j}\Psi\left(\alpha_. + N_i\right)$$
$$-\frac{d^2}{d^2\alpha_0}\Psi\left(\alpha_. + N_i\right) + \frac{d^2}{d^2\alpha_0}\Psi(\alpha_0) + \frac{d^2}{d\alpha_0 d\alpha_{j'}}\Psi\left(\alpha_. + N_i\right).$$

These moments will be used to obtain the empirical Bayes estimators of $\boldsymbol{\beta}$.

4 Estimation

Under the squared error loss function, the mean of the posterior distribution is the Bayes estimator for $\boldsymbol{\eta}_i$. That is, as given in (3-2),

$$\eta_{ij}^B = E(\eta_{ij}|\mathbf{y}_i,\boldsymbol{\alpha}).$$

To obtain the empirical Bayes estimator, we need to replace $\boldsymbol{\alpha}$ by its estimate $\mathbf{r}$. Using the method of moments, the estimate $\mathbf{r}$ is obtained. Details of this procedure are given in Eskandari and Meshkani (2000). Hence

$$\eta_{ij}^{EB} = E(\eta_{ij}|\mathbf{y}_i, r) \equiv f(r_j),$$

where

$$r_j = \frac{\bar{M}_j\bar{X}_j}{\bar{M}_j(1-\bar{M}_j) - \bar{X}_j},$$

with $\bar{M}_j = \frac{1}{I}\sum_{i=1}^{I}(\frac{y_{ij}}{N_i})$ and $\bar{X}_j = \frac{1}{I}\sum_{i=1}^{I}(\frac{y_{ij}}{N_i}(1-\frac{y_{ij}}{N_i}))$.
To build a relation between η_{ij}^B and η_{ij}^{EB}, we use the Taylor-series expansion

$$f(r_j) = f(\alpha_j) + \left.\frac{\partial f(r_j)}{\partial r_j}\right|_{r_j=\alpha_j}(r_j - \alpha_j) + \cdots + \left.\frac{\partial^k f(r_j)}{\partial r_j^k}\right|_{r_j=\alpha_j}\frac{(r_j-\alpha_j)^k}{k!} + \cdots$$

Since $\Psi(x) \sim log(x) + \frac{1}{2x}$ (Harisson et al., 1985), from (3-2)

$$f(r_j) = f(\alpha_j) + o(\frac{1}{k}).$$

Therfore, we can write

$$\eta_{ij}^{EB} = E(\eta_{ij}|\mathbf{y}_i, \boldsymbol{\alpha}) + e_{ij} \qquad (4-1)$$

where $E(e_{ij}) = 0$,

$$Var(e_{ij}) = Var(\eta_{ij}^{EB}|\mathbf{y}_i) = \frac{d}{d\alpha_j}\Psi(\alpha_j + y_{ij}) + \frac{d}{d\alpha_0}\Psi(\alpha_0 + y_{i0}) = \sigma_{ji}^2,$$

and $Cov(e_{ij}, e_{i'j}) = 0$. This is because for each j, the row elements e_{ij} correspond to independent replicates of the observations.

Now suppose for $j = 1, 2, \cdots, K$, $\boldsymbol{\beta}_j^{'} = (\beta_{j0}, \beta_{j1}, \cdots, \beta_{jq})$ is the vector of regression parameters coressponding to the vector of q covariates $\mathbf{x}_i^{'} = (1, x_{i1}, \cdots, x_{iq})$. The assumed structure for the logistic regression is

$$H_0 : \eta_{ij} = \mathbf{x}_i^{'}\boldsymbol{\beta}_j. \qquad (4-2)$$

Furthermore, from (4-2), for $i = 1, 2, \cdots, I$, we can write

$$H_0 : \boldsymbol{\eta}_j = \begin{pmatrix} \eta_{1j} \\ \vdots \\ \eta_{Ij} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^{'} \\ \vdots \\ \mathbf{x}_I^{'} \end{pmatrix} \boldsymbol{\beta}_j = X\boldsymbol{\beta}_j,$$

Thus (4-1) is rewritten as

$$\boldsymbol{\eta}_j^{EB} = \begin{pmatrix} \eta_{1j}^{EB} \\ \vdots \\ \eta_{Ij}^{EB} \end{pmatrix} = \begin{pmatrix} E(\eta_{1j}|\mathbf{y}_1) \\ \vdots \\ E(\eta_{Ij}|\mathbf{y}_I) \end{pmatrix} + \begin{pmatrix} e_{1j} \\ \vdots \\ e_{Ij} \end{pmatrix} = E(\boldsymbol{\eta}_j|\mathbf{Y}) + \mathbf{e}_j$$

or in matrix form as

$$\boldsymbol{\eta}_j^{EB} = XE(\boldsymbol{\beta}_j|\mathbf{Y}) + \mathbf{e}_j,$$

where

$$\mathbf{e}_j \sim [0, \Sigma_j], \qquad (4-3)$$

and $\Sigma_j = Diag\{\sigma_{j1}^2, \sigma_{j2}^2, \cdots, \sigma_{jI}^2\}$. Thus, a weighted least squares estimate of $E(\boldsymbol{\beta}_j|\mathbf{Y})$ is

$$\Big(E(\boldsymbol{\beta}_j|\mathbf{Y})\Big)^{EB} = \Big(X'\Phi_j X\Big)^{-1}\Big(X'\Phi_j\boldsymbol{\eta}_j^{EB}\Big), \qquad (4-4)$$

where $\Phi_j = \Sigma_j^{-1}$.

The estimate $\Big(E(\boldsymbol{\beta}_j|\mathbf{Y})\Big)^{EB}$ is considered as an empirical Bayes estimate of $\boldsymbol{\beta}_j$ required in (4-2), which in turn provides the estimates of $\boldsymbol{\eta}_j$. This finally leads

to the estimates of $\mathbf{P}$ via (2-1). Hence, the proposed model gets estimated. The covariance matrix of $\left(E(\beta_j|\mathbf{Y})\right)^{EB}$ is

$$Var\left[\left(E(\beta_j|\mathbf{Y})\right)^{EB}\right] = \left(X'\Phi_j X\right)^{-1} X'\Phi_j \Sigma_j \Phi_j X \left(X'\Phi_j X\right)^{-1}$$
$$= \left(X'\Phi_j X\right)^{-1}. \qquad (4-5)$$

5 Example: Analysis of Pregnancy Data

In this Section, we analyze a subset of the pregnancy outcome in consanguineous marriages. Schull (1958), analyzed these data using a frequentist approach, and Forster (1999) reanalyzed it via Metropolis-Hastings algorithms.
The study sample, according to degree of consanguinity between the parents, included 6258 pregnants women in three districts of Shizuka city, Japan. Here, we have two covariates, R≡Residence(Rural district, Intermediate district, Urban district) and C≡Consanguinity(no ralation , 2nd cousions, 1st cousions). The categories of Death are A≡Abortion, S≡ Stillbirth, U≡ in less than 12, V≡ in 13-60 and Su≡ Survived. We consider multinomial regression model (4-1) for pregnancy outcome, with district as a categorical covariates. The design matrix X is

$$X = \begin{pmatrix} 1 & -1 & -1 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

With the category labelling in Schull, the baseline category for model (4-1) is survived. Table 1 present the estimated superparameters and statistics required for each cells. Furthermore we can estimate, $\frac{d}{d\alpha_j}\Psi(\alpha_j + y_{ij})$ matrix for $j = 0, 1, 2, 3, 4$, hence by (3-3) the estimate covariance matrix are shown in Table 2. From Tables 1 and 2, we can obtain estimates of the parameters and hence the estimated logits for any pair of Death groups for each cell. These results are shown in Table 3. For instance, the second equation is

$$Ln\left(\frac{\hat{P}_S}{\hat{P}_{Su}}\right) = -2.08 - 0.735R + 0.2918C$$

Using this result, for residence, the estimated odds of Stillbirth instead of Survived are $exp(0.735) = 2.08$ times higher for Rural district than Intermediate district and

$(exp(0.735) = 2.08)/(exp(-0.735) = 0.479) = 4.35$ times higher for Rural district than Urban district, and for Consanguinity, the estimated odds Stillbirth instead of Survived are $exp(-0.2918) = 0.75$ times higher for no relation than 1st cousins and $(exp(-0.2918) = 2.08)/(exp(0.2918) = 1.338) = 0.56$ times higher for no relation than 2nd cousins. Finally, Table 4 reports expected probabilities for the logistic regression models. In column survived, given Residence is Urban district and Consanguinity is no relation, Muximum probability is 0.92, and in colomn stillbirth from top to blow, we can obtain expected numbers of pregnancy by multiplying each probability by the number of observations at that Residence and Consanguinity level.

Table 1: The Estimated Superparameters and Statistics Requaired

$Death$	$Abortion$	$Stillbirth$	$\leq 12Months$	$13-60Months$	$Survived$
M_1	.028	.0156	.0595	.0261	.871
X_1	.0272	.0154	.05595	.0254	.1123
M_2	.0063	.0063	.081	.0375	.87
X_2	.00626	.00626	.074	.0361	.1131
M_3	.042	.011	.069	.0363	.841
X_3	.040	.0108	.0642	.0349	.1337
M_4	.0251	.0075	.048	.0285	.891
X_4	.024	.0074	.0456	.0277	.0971
M_5	.0325	.0029	.074	.0296	.8609
X_5	.0314	.0289	.0685	.0287	.1198
M_6	.0381	.011	.061	.044	.846
X_6	.0366	.01088	.0573	.0421	.1303
M_7	.0129	.0092	.0386	.0258	.913
X_7	.0127	.00912	.0371	.0251	.0794
M_8	.057	0	.0143	.0286	.9
X_8	.054	0	.01409	.0278	.09
M_9	.037	.0037	.0741	.048	.837
X_9	.0356	.00368	.0686	.0457	.136
$\bar{M}$	.031	.0078	.0577	.0338	.8699
$\bar{X}$	.0297	.0074	.054	.0326	.1124
$\mathbf{r}$	2.888	.868	7.061	19.162	126.343

Table 2: The Estimated Covariance Matrix

ij	σ^2_{4i}	σ^2_{3i}	σ^2_{2i}	σ^2_{1i}
1	.0329	.0610	.01549	.02238
2	.22413	.39204	.0464	.0389
3	.0543	.1843	.0307	.0306
4	.0142	.0468	.0074	.0105
5	.0694	.3920	.0307	.0337
6	.0267	.0878	.0162	.0170
7	.096	.1559	.0350	.0297
8	.1346	.4884	.1164	.046
9	.074	.392	.0363	.0306

Table 3: Estimated Parameters in logit Models for Pregnancy Data using Survived as baseline category

Death.	*Source*	$\hat{\beta}$	$Var(\hat{\beta})$	Z	*Sig.*
Abor.	*Cons.*	-3.49	.009	-36.78	0
	Resid.	.0021	.005	.03	.95
	Consan.	.249	.003	4.54	0
Still.	*Cons.*	-2.08	.008	-23.18	0
	Resid.	$-.735$	.006	-9.83	0
	Consan.	.2918	.003	5.23	0
$\leq 12 Mon.$	*Cons.*	-4.236	.041	21.04	0
	Resid.	$-.363$	.031	-2.076	.02
	Consan.	$-.014$	.018	$-.107$	.67
$13-60 Mon.$	*Cons.*	-3.5	.017	26.68	0
	Resid	$-.32$	.012	-2.89	.003
	Consan.	.37	.006	4.87	0

Table 3: The Expected Probabilities for each cell

Resid.	*Consan.*	*Abortion*	*Stillbirth*	$\leq 12 Mon.$	$13-60 Mon.$	*Survived*
Rural	*no relation*	.026	.012	.109	.026	.827
	2nd cousins	.038	.012	.145	.034	.771
	1st cousins	.053	.012	.188	.0419	.705
Interm.	*no relation*	.019	.009	.053	.027	.891
	2nd cousins	.027	.009	.069	.034	.861
	1st cousins	.038	.008	.091	.042	.820
Urban	*no relation*	.014	.006	.026	.028	.925
	2nd cousins	.021	.006	.035	.035	.903
	1st cousins	.028	.006	.043	.042	.881

References

Agresti, A. (1990). *Categorical Data Analysis*. New York, John Wiley.

Bickel, J. and Doksum, A. (1977). *Mathematical Statistics : Basic Ideas and Selected Topics*. Holden-Day, San Francisco.

Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler. *The American statistician*, **46**, 167-174.

Dellaportas P. and Forster J.J. (1999). Markov chain Monte Carlo Model Determination for Hierarchical and Graphical Log-Linear Models. *Biometrika*, **86**, 615-633.

Eskandari, F. and Meshkani, M. R. (2000). Empirical Bayes analysis of log-linear models for generalized finite stationary Markov chain. *proceeding of the Fifth Iranian Statistics Conference*. Isfahan University of Technology Isfahan.

Forster, J. J. (1999). Markov Chain Monte Carlo Exact Inference for Binomial and Multinomial Logistic Regression Models. *Technical Report*. University of Southampton.

Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97-109.

Kendall, M. and Stuart, A. (1977). *The Advancved Theory of Statistics*.Vol. **1**. London, Griffin.

Laird, N. M. (1978). Empirical Bayes Methods for two-way contingency tables. *Biometrika* , **65**,581-590.

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

Leonard, T. (1975). Bayesian Estimation Methods for Two-Way Contingency Tables. *Jour. R. Statist. Soc.*. **B, 37**,23-37.

Meshkani, M. R. and Billard, L. (1992). Empirical Bayes Estimators for a Finite Markov Chain. *Biometrika*. **79**,**1**, 185-93.

Nazaret, W. A. (1987). Bayesian Log Linear Estimates for Three-Way Contingency Tables. *Biometrika*. **74**,**2**, 401-10.

Newton , M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Jour. Roy. Statist. Soc.* B, **56**, 3-48.

Ntzoufras, I. (2000). Byesian Modelling of Outstanding Liabilities Incorporating Claim Count Uncertainty. *Technical Report*. Athens University of Economics and Business.

Ntzoufras, I. (1999). Gibbs variable selection using BUGS. *Tecnical Report*, Dep. of stat., Athens University.

Ntzoufras, I., Dellaportas, P. and Forster, J.J. (1999). Specification and interpretation of prior distributions for variable selection in linear model. *Proceedings of the Fourth Hellenic-European Conference on Computer Mathematics and its Applications* (E. A. Lipitakis, ed.).

O'Hagan, A. (1995). Kendall's Advanced Theory of Statistics **2B**, *Bayesian Inference*. Edvard Arnold, London.

Raftery, A. E. (1996). Bayesian Model Selection in Social Research. *Sociological Methodology* (P.V. Marsden ed.). Oxford: Blackwell.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. second edition. New York, John Wiley.

Schmidt, P., and Strauss, R. (1975). The prediction of occuption using multiple logit models. *Internat. Econ. Rev.* **16**, 471-486.

Schull, W. J. (1958). Empirical risks in consanguineous marriages: sex ratio, malformation and viability. *Amer. Jour. of Human Genetics*, **30**, 294-212.

Stochastic Models for the Planning of Pharmaceutical Research

Gittins, J.

P17013

Unieversity of Oxford, UK.

Abstract. Recent work on two aspects of this large topic are described. Both are concerned with the 'research' or 'discovery' phase of pharmaceutical R & D, during which there is a search for new chemical entities which are sufficiently promising to be used in clinical trials.

The first aspect is a statistical procedure for the selection of compounds to be submitted to the screening tests which characterise this search process. Secondly, a stochastic optimisation model for the allocation of resources over the successive stages of a discovery-phase project is described.

1 GENERAL DESCRIPTION

1.1 Summary

CPSDAI is a computer program to help chemists to use the results of testing compounds for some form of desirable activity as a guide to finding highly active compounds. It is based on statistical principles. The program is written in Fortran 77 and may be run on any platform with access to at least 1Mb of memory.

1.2 Purpose

In pharmaceutical and agro-chemical research thousands of chemical compounds are synthesised and screened for every new product which is ultimately marketed. In the early stages of screening one of the main aims is to identify compounds with a sufficiently high level of interesting activity to warrant more extensive testing. Targets of this kind may be defined in terms of any appropriate scale for measuring activity - for example the percentage of test animals showing signs of improvement at a given dosage. The record of activity measurements (or *scores*) for successive compounds gives some indication of the attainability of such a target, and hence either provides encouragement to carry on selecting new compounds for screening with approximately the same characteristics as before, or indicates that some change of direction is necessary if the target is to be achieved within a reasonable time. *CPSDAI* analyses the past record along these lines. It may be used in the context of high-throughput screening or in following up known lead compounds. Preliminary versions have been around for several years and have been used in the pharmaceutical and agro-chemical industries and in the design of abrasive-resistant materials. A flexible user-friendly version is now available. Readers wishing to try it out should contact the author.

1.3 Outputs

CPS stands for *Current Probability of Success*. This is the estimated proportion of compounds with scores above the target, assuming that the distribution of future scores is essentially the same as the distribution of past scores. The case of greatest interest is often for a target which is higher than the largest score so far achieved. The CPS is then estimated by extrapolating the distribution of past scores beyond the target in such a fashion as to minimise the inevitable risk of error in such a procedure.

The compounds which have been tested can often be divided into different groups, defined so that the compounds within a group have similar chemical structures. A different CPS can then be calculated for each group. In general a group with a high CPS is a better prospect for achieving the target by screening more compounds than is a group with a low CPS. However, for two groups with similar values for the CPS preference should be given to the group for which the number of compounds so far tested is least. This is because when only a few compounds have been tested there is a good chance that the CPS substantially underestimates the proportion of compounds which in the long run would turn out to have scores above the target, and some priority should be given to resolving this uncertainty, so that in future compounds may be selected for testing in a more informed way.

The *Dynamic Allocation Index (DAI)* for the scores achieved by a set of compounds belonging to a particular group is greater than the CPS for the same compounds by an amount which measures the importance of reducing uncertainty about the group as a whole. It is calculated so that, if there were several distinct groups of compounds available for screening, and the compounds were to be screened one by one, the expected number of compounds which would have to be screened before attaining the target score would be minimised by at each stage allocating for screening a compound from the group with the largest current value of the DAI. Thus the DAI is an index of priority. It is called *dynamic* because it changes whenever the score for an additional compound is added to the distribution of scores on which it is based. Of course compounds are not usually screened one at a time, but the DAI remains a useful indicator.

In addition to the CPS and the DAI for each specified group of compounds the $CPSDAI$ program calculates upper and lower 90% probability limits, and the median value, for the number of additional compounds from the group which would need to be screened before finding a compound which attains the target. The scores for the different groups are summarised by means of histograms.

1.4 Inputs

The inputs required by the program are: any assessments of the likely distribution of scores within each group that the user would like to take into account to supplement the information yielded by the scores themselves; the actual scores; and an estimate of the experimental variation to which the observed scores are subject.

1.5 Validity

The outputs are valid if it is a reasonable assumption that the scores of the as yet untested compounds in a group are randomly selected from the same distribution

of scores as those of the compounds which have already been tested. For this to be true it is not necessary that the structures of the compounds to be screened should be determined in a random manner, which is not often true. Care should, however, be taken to ensure that the set of already tested compounds on which an analysis is based do cover a range of structures which is broadly similar to the range from which further compounds are likely to be drawn. For example, if it is proposed to synthesise a series of derivatives of a given lead compound which all involve modifications at a particular site, whereas all previously tested derivatives of that lead compound have been at other sites, the only directly relevant previously tested compound may well be the lead compound itself. Later on the search may range once again over a wider class of derivatives, and the appropriate reference set of tested compounds would then widen again to include at least some of the compounds which were excluded when the focus was on modifications at a particular site.

2 A LITTLE MORE DETAIL

2.1 The General Idea

When there are a number of different jobs to be done, projects in which we might invest, or lines of research we might pursue, the question arises of how we should assign priorities so as to minimise costs or maximise rewards. A *dynamic allocation index (DAI)* is a number associated with any particular alternative, with the property that the optimal policy is to assign priority to the alternative with the largest DAI. These indices typically change as work progresses, so that an optimal policy may well switch back and forth between projects. DAIs with these properties may be defined for a variety of probabilistic models, some of which are relevant to aspects of chemical research. These are reviewed in a book on the subject (Gittins, 1989).

One such model has been designed as an aid in the selection of formulations for screening in new-product chemical research. The idea is that within a typical project there are a number of alternative *routes* (or groups of compounds or other formulations) representing different possible lines of attack, which vary in difficulty and which may lead toward the solution of the chemical problem for which the project was set up. $CPSDAI$ makes calculations for each of these routes. The different routes are defined by the different classes of formulations that could be tested. These might be suggested by various chemical hypotheses as to the ways in which the desired result might be achieved, or simply emerge empirically by noting those formulations which have already been found to be reasonably promising. They might, for example, correspond to clusters identified by some form of statistical cluster analysis, perhaps in association with a consideration of the physical properties of the molecules. Sometimes, as in the example illustrated in figure 1, there is only one route.

We suppose that for each formulation a score may be calculated from the test results, the more promising formulations being those with the higher scores. When compounds are being tested to find one that would be suitable for use as a drug, the important thing is the level of therapeutic activity. This is frequently measured by a single number, such as the proportion of diseased animals that recover when

treated with the compound, which we may regard as the score for the compound. In this respect, however, pharmaceutical research is the exception rather than the rule. Generally speaking, several different attributes are relevant to the desirability or otherwise of a formulation. However, provided that it is possible to give an order of preference to formulations on the basis of the values taken by the attributes, a single score for each formulation may fortunately still be determined. We shall assume that this has been done, without wishing to suggest that this is always a simple task.

Let T be a *target* score, whose achievement would represent a significant step forward in a project. The choice of T is at the chemist's discretion, but it should be large enough so that a formulation with a score above T is worthy of serious further consideration, and not so large that the project is likely to have to be terminated before such a formulation is found.

A histogram of the scores of the formulations that have so far been tested from a route, showing also the value of T, gives a good indication of the promise of the route, as figure 1 illustrates. The figure shows the successive histograms for a hypothetical route after 2, 8 and 16 formulations have been tested, respectively. Two possible targets, T_1 and T_2, are also shown, of which T_2 is the more ambitious.

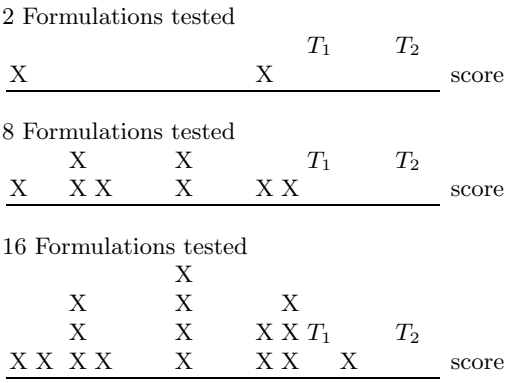

Fig. 1. Successive histograms for a route

Not much can be said after just two scores have been obtained, although they are sufficiently widely spread to give grounds for hoping that the targets may both be attainable. When eight formulations have been tested the picture is clearer. The suggestion that T_1 is likely to be reached fairly soon is strengthened, but it is beginning to look as though T_2 may not be reached for a long time, unless the chemist hits on some method of finding formulations with higher scores. After testing 16 formulations this picture is confirmed.

The dynamic allocation index for a route is a number that quantifies these impressions obtained from the histogram of scores so far achieved. It is a measure of the current promise of a route, insofar as this is reflected in the scores of the

formulations that have already been tested. In many cases chemists have additional information, perhaps sometimes amounting to little more than a hunch, which leads them to believe that a route is either more or less promising than the DAI suggests. When this happens they will, quite rightly, take such considerations into account in deciding on which routes to concentrate, and availability and ease of synthesis are, of course, also relevant. What the $CPSDAI$ calculations for the various routes do is to provide an aid to the continual dialogue between the chemist and the experimental data, and to indicate those routes which, on the basis of past results, and taking account of any prior beliefs, it seems most profitable to pursue.

A DAI which prioritises routes on the basis of minimum expected cost to reach the target may be obtained by dividing the DAI given by $CPSDAI$ by the cost of testing a compound from a given route. Alternatively, to minimise expected time we may divide by the time taken to test a compound.

It is worth noting that the appropriate choice of routes depends on the level of the target. This is illustrated by the route whose history is shown in figure 1. It is fairly obvious that for the target T_2 the DAI for the route must decrease at each successive stage, as it becomes increasingly clear that most of the scores are well below T_2. For T_1, on the other hand, the DAI may well increase.

This phenomenon is not surprising. It is a reflection of the fact that if what is required is a modest improvement over current performance levels, it is probably best to try modifications of one of the currently used formulations; whereas if the target is a really substantial improvement, it is worth considering completely different and relatively untried routes. It does, however, show the importance of an appropriate choice of target. There are occasions when it is worthwhile applying $CPSDAI$ simultaneously for two or more different target values.

2.2 Some Technicalities

For a given route a score of I is chosen, either by the program or by the user, and the scores are rescaled linearly so that $I = 0$ and $T = 1$. The number of transformed scores above any positive value x is assumed to be proportional to $exp(-\theta((x+2)^r - 2^r))$ for appropriately chosen parameters θ and r. The value of I needs to be close enough to T for a model of this form to be a reasonable fit to the distribution of scores, but not so close that the frequency of scores above I is very low. Formulations with scores above I are termed *interesting* formulations. Let p be the proportion of formulations with scores below I. These *uninteresting* formulations are essentially modelled only in terms of estimates for p, without taking further account of their distribution.

The DAI for a route is equal to the posterior probability, given the results from the route to date, that the next formulation tested has a score exceeding T, which we term the *current probability of success (CPS)* , together with an upward correction which is large if the uncertainty associated with that posterior probability is also large. Since our primary interest is in routes for which at most one score over T has so far been noted, it follows that the CPS is something like an extrapolated estimate of the tail area of the distribution of scores in the route, based on the current histogram of scores. This means that both DAI and CPS are subject to considerable error if there is any appreciable departure from the assumed form for the distribution of scores, and any such error will be particularly great if the

extrapolated distribution is fitted to those parts of the histogram for which the scores are well below T. This is the reason for treating scores differently according as they are above or below I. In fact the treatment of scores near I allows a gradual transition from uninteresting scores to interesting scores, so the procedure is not sensitive to whether particular scores are just over or just under I.

On the basis of information provided by the user the program sets up prior distributions for p, r and θ. These will have large or small variances depending on the strength of the user's prior information. A beta prior distribution is assumed for p, and gamma distributions for r and θ.

For example, the prior density function for r is of the form $\Gamma(n)^{-1}\Sigma^n x^{n-1}exp(-\Sigma x)$, where $x = r$. Values of r which are much greater than one are rather implausible - for example a normal distribution of scores would mean $r = 2.0$. Reasonable values for many purposes are $n = 2.0$ and $\Sigma = 2.0$. With these values the prior probability that r lies between 0.1 and 3 is 0.965. The program sets $n = \Sigma$, with a common value of at least 2.0, and as close to 2.0 as is consistent with the input prior probabilities.

The scores of the formulations which have been tested are used to modify the prior distributions for the parameters by means of Bayes theorem. The program then uses the resulting posterior distributions to calculate CPS, DAI, and a prediction interval for the number of further formulations which will need to be tested to reach the target. All these are defined in terms of the true score, without experimental error. The assumption is that if the observed score of a formulation is near the target the formulation is likely to be further investigated, at least to some extent, so compounds which achieve the target are unlikely to be missed.

Experimental error is allowed for by assuming the differences between x^r and y^r, where y is the observed score and x is the true score, to be normally distributed, with a variance calculated from user-provided input.

The calculation of DAI is based on tables given in Gittins and Jones (1974) and Gittins (1989). Some approximation is needed as these tables are for the case when r is known to be equal to one. An account of these calculations, and of the underlying theory, is given in Gittins (1994).

2.3 Some Advice on Using $CPSDAI$

The $CPSDAI$ calculations for a route are on the basis that the scores achieved by the different chemical formulations in the route may be regarded as independently drawn from the distribution of scores for the entire set of chemical formulations defining the route. Now this is a considerable oversimplification. The point is that formulations which are similar in chemical composition are also likely to have similar scores, and there are certain to be degrees of similarity in chemical composition even for formulations belonging to the same route.

A basic organic molecule, for example, may have one or more sites at which side-chains could be attached. Compounds all having this same basic component may then be compared on the basis of whether there are side-chains, and if there are whether they have common features, such as halogen atoms. Then there are those formulations which consist of mixtures of compounds, a class which includes, for example, nearly all detergents. There will clearly be a tendency for those mixtures

which contain the same, or similar, ingredients to have similar scores. The homely term *ingredient*, which is natural when thinking of mixtures of compounds, we shall also use in reference to particular features of complex molecules.

All this means that the total amount of variability of the scores in a route should be recognised as being attributable to a number of specific identifiable causes. To the statistician this naturally suggests what is known as a *components of variance* model: a model, that is to say, in which the over-all variation of scores within a route is divided into components assigned to the different individual ingredients of which the formulations in the route may consist. A model of this kind would, in principle, lead to a more general $CPSDAI$ analysis, indicating which ingredients should be changed as well as the promise of the route as a whole. There are, however, some difficulties, the most important being that the assumptions of the model would probably not hold. It is often impossible to specify in advance what types of ingredient might be used, and, contrary to the model, the different ingredients would not typically contribute independent amounts to the score. In addition, the computations required would be quite unmanageable, and the resulting management aid would in any case be distinctly unwieldy. Some alternative, and less formal, method of coping with the problem of different types of ingredient is called for.

One point which needs to be remembered is that the formulations tested are selected by a chemist from the total population of possible formulations in a route as being worth testing. They would be wise to try at an early stage those types of ingredient which are particularly promising. It would be a waste of time to test a large number of formulations which differ only in respect of types of ingredient which do not cause much variation in score. Thus there should be a tendency for the important types of variability all to be explored at an early stage, a consideration which favours the validity of our simplified model, up to the point at which the search starts to be concentrated on sub-populations of particular interest. Nonetheless, for the sake of building up chemical knowledge in a systematic fashion the chemist may well test more formulations of a fairly similar nature than is justified simply on the basis of a rapid exploration of all the important sources of variability. This leads to the idea of limiting the number of formulations whose scores are used as a basis for $CPSDAI$ calculations according to the number of distinct ingredients which have been used. With these considerations in mind, the following scheme for applying $CPSDAI$ is suggested:

As a first essential, the score for every formulation tested in a route should be plotted sequentially, together with the CPS and the DAI. Histograms of scores should also be plotted for each route and regularly updated.

In general, every formulation tested should be counted in the calculations, except when the same ingredients are mixed together in different combinations or proportions. When this happens the number of scores which are counted from a group of chemicals with ingredients in common is limited to the number of ingredients in the group which occur just once, plus twice the number of ingredients which occur more than once. The score of any formulation which includes an ingredient which does not occur anywhere else is counted. If more formulations from the group have been tested than the number of scores which are allowed to count, then the scoring formulations are selected so that (i) as far as possible each ingredient is in a scoring formulation equally often, and (ii) subject to (i) the scoring formulations

are chosen so as to be representative of the range of variation of the scores recorded for the group. If some particular class of ingredient has members which differ only in hydrocarbon chain length, then only three ingredients belonging to the class are counted as separate ingredients for the purposes of this paragraph: those ingredients having the greatest and the smallest chain lengths, and one with an intermediate chain length.

There will be occasions when a particular line of investigation becomes so promising and permanent as in effect to constitute a new route. It may be obvious on chemical grounds that this is happening, or it may be suggested in the first instance by a change in the general level of scores. When this happens, those formulations which belonged to the old route but not to the new one must be excluded from future calculations.

The notion of different particular lines of investigation all within the same route suggests a further application of $CPSDAI$ which might sometimes be useful. This is to keep a record of the highest score obtained by a formulation within each line of investigation and to carry out an analysis on this basis. This would give an indication of the number of similar lines of investigation required before one is found which produces a formulation with a score above the target. Its usefulness will depend on the number of potential sub-routes being large.

2.4 An Example

Target	125	125
Cutoff point for interesting formulations	50	70
Dynamic allocation index	0.0036	0.0020
Probability next formulation reaches the target	0.0028	0.0013
Lower 90% limit for the number of further formulations required to reach the target	19	42
Median for the number of further formulations required to reach the target	380	1400
Upper 90% limit for the number of further formulations required to reach the target	5900	96000

Table 1. Results of analysis for route 1.

In tables 1 and 2 the results of tests on formulations drawn from two routes in a research program designed to produce a herbicide are summarised. The raw data for each formulation consisted of assessments of the severity of the effect when the formulation was applied to a particular plant species in a particular way and at a given dosage level. In all, forty observations were available for each formulation for different combinations of these factors. The target of 125 represents an existing herbicide, the aim being to find a formulation which is at least as toxic to the relevant plant species. Similar raw data were available for this target herbicide.

Target	125	125
Cutoff point for interesting formulations	50	70
Dynamic allocation index	0.0028	0.0012
Probability next formulation reaches the target	0.0021	0.00063
Lower 90% limit for the number of further formulations required to reach the target	24	87
Median for the number of further formulations required to reach the target	410	2300
Upper 90% limit for the number of further formulations required to reach the target	3700	67000

Table 2. Results of analysis for route 2.

Scores for each formulation were obtained by first taking an appropriate weighted average of the differences between the forty observations for target herbicide and formulation respectively, and then adding a constant so that the lowest observed score was zero. The scores for each route are summarised in figures 2 and 3.

For both routes the reason for the relatively large number of formulations with scores near the maximum is because several formulations all with very similar molecular structures were tested. The next step was to remove some of these from the analysis, along the lines mentioned in the previous section. The resulting reductions are shown by the dotted lines in figures 2 and 3. The analysis given in tables 1 and 2 was based on the resulting modified sets of scores.

The choice of the cutoff point between uninteresting and interesting formulations, not surprisingly, has a significant influence on the analysis. The overall message from this data is, however, fairly clear. For both routes there is a reasonable chance that a few hundred more formulations sampled along similar lines would produce one which achieved the target. However the number needed might well also run into thousands, or even tens of thousands, without some further refinement of the populations of formulations being sampled.

PartI

Algorithms for Allocating Resources to Multi-Stage Pharmaceutical Research

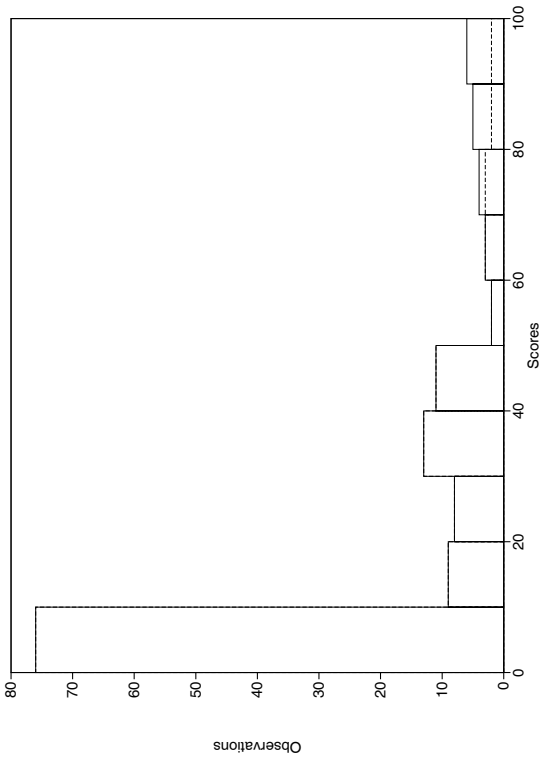

Fig. 2. Data for route 1 (dots indicate reduced numbers of observations).

Projects

3 Summary

Two alternative stochastic models are described for a stage in a multi-stage pharmaceutical research project. These may be used to estimate the most profitable number of scientists to allocate. General properties of these optimal allocations are derived and algorithms to calculate the optimal allocations are described, together with examples of their use. Optimal allocations are higher than those which are likely to complete the project in the minimum number of scientist-years.

4 Introduction

The process which begins with exploratory research in an industrial pharmaceutical laboratory and culminates in the marketing of a new drug is usually described either as *research* (or as *discovery*) in its early stages and later on as *development*. The dividing point between research and development is often defined to be when a

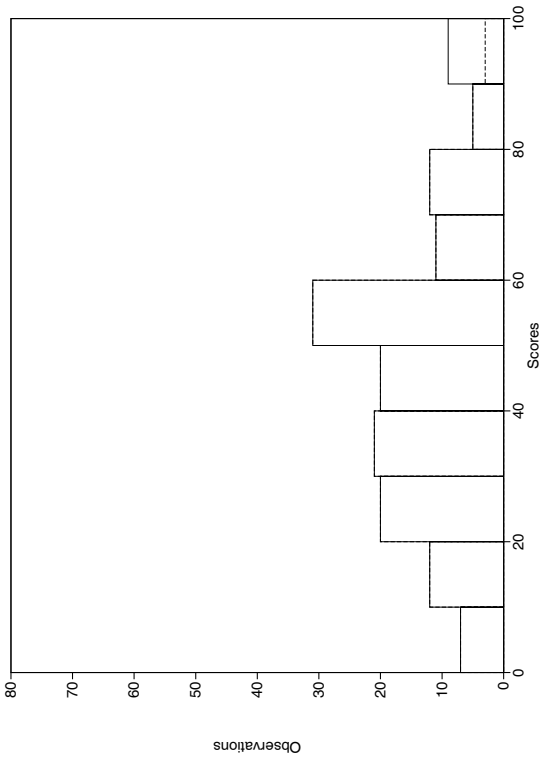

Fig. 3. Analysis of route 2 (dots indicate reduced numbers of observations).

compound is designated as a potential new drug, or *development compound* often referred to as a *new chemical entity*. At, or soon after, this point the compound moves out of the laboratory and clinical trials begin. On average fewer than 20% of development compounds finally emerge as marketable drugs, and consequently more than one development compound are usually selected from any given project. The research phase of a project is characterised by screening tests on a large number of compounds, many of which may have been synthesised for the purpose. The duration of the research phase varies widely, ten years being typical. The timescales in development are more predictable, eight years being typical.

For the research to be profitable it is important to keep these long periods within bounds. Gittins (1997) investigates the relationships between profitability and the numbers of scientists allocated at the different stages of the research phase of a project, using a stochastic model. The general conclusion is that larger project teams than those which are typical of current practice would in some cases be much more profitable.

Here the stochastic model is described and used to derive some of the properties of an optimal policy. These in turn have been used to construct algorithms for the calculation of optimal policies for two versions of the model, and examples are

given of these policies. More details of the structure of the process of research in the pharmaceutical industry are given, for example, by Boschi (1982), Bergman and Gittins (1985), Spilker (1989), and in reports by Andersen Consulting (1998) and by Price Waterhouse Coopers (1998).

5 Modelling Assumptions

The research phase of a project consists of four stages: stage 1, preliminary; stage 2, identifying the first lead compound; stage 3, identifying the first development compound; and stage 4, identifying a further development compound. The stages take place in sequence, and stage 4 may be repeated (or omitted) as often as is necessary to produce the required number of development compounds.

If u_i scientists are allocated to stage i of a project the rate of progress is $e_i(u_i)$. If the rate of progress was simply proportional to the number of scientists, we could choose appropriate units so that $e_i(u_i) = u_i$. However it is widely believed that there is an optimal size for a team of scientists, and that higher or lower numbers lead to a loss of efficiency. To model this we use an effectiveness function e (dropping the subscript i) such that

$$max_u(e(u)/u) = 1.$$

Let u_{eff} be the value of u such that

$$e(u_{eff})/u_{eff} = 1.$$

Thus $e(u)/u$ is the relative efficiency of a team of u scientists compared with u_{eff}, the most efficient team size. The numerical calculations used functions of the form

$$e(u) = au^2/(1 + bu^r).$$

These functions mean that the relative efficiency has the required unimodal shape as a function of u.

Two alternative models for a research stage will be described, the second model being more detailed and realistic, as well as leading to more complicated calculations.

Model 1

When u scientists are allocated, the time T needed to complete the stage is $X/e(u)$, where X is known in advance, and is the number of scientist-years needed to complete the stage when $u = u_{eff}$. On completion of the stage it has either been successful, with probability p, or unsuccessful, with probability $1 - p$. At stage 4, $p = 1$. After an unsuccessful stage the project is terminated.

Model 2

The number of scientists allocated at a time t after the beginning of the stage is $u(t)$. Thus we may define the *effective* work done up to time t to be

$$x(t) = \int_0^t e(u(s))ds.$$

The effective work X required to complete the stage successfully has distribution function F and density function f, and the time required is T, so that $X = x(T)$. It follows that T has the distribution function $F(x(\cdot))$ and density function $e(u(\cdot))f(x(\cdot))$.

The expected cost of the project is measured in scientist-years, on the basis that the cost of a scientist includes the cost of overheads, equipment, accomodation, and technical and secretarial assistance. Future costs are discounted by a factor which expresses the lower value of a sum of money in the future compared with the value of the same sum if it was available immediately. If the present value of *£1* which becomes available (indexed for inflation) after t years is $£exp(-\gamma t)$, the cost in scientist-years of employing u scientists for t years is

$$\int_0^t u\ exp(-\gamma s)ds = u\gamma^{-1}(1 - exp(-\gamma t)). \tag{1}$$

Obsolescence means that the exponential rate γ_1 at which future rewards from the project are discounted is higher than the rate γ which applies to future costs.

The expected value of a development compound available now for clinical trials is $£V$. This expected value is based on the distribution of possible cash flows resulting from a new drug, and takes account of all costs, and the possibility that the compound may not survive clinical trials. When the possibility that any of the first three stages of research may be unsuccessful is taken into account, and the value of the development compound is discounted to allow for the fact that it is not available now but after a time $T_1 + T_2 + T_3$, the expected value of the first development compound which may emerge becomes

$$p_1p_2p_3exp[-\gamma_1(T_1 + T_2 + T_3)]V$$

for Model 1.

Subsequent development compounds have a lower expected value for two reasons. First there is an additional discount factor because any rewards occur later. Secondly, the expected market for each successive compound in the series is reduced by competition from its predecessors. The effect of this competition is to multiply the value of the j'th compound in the series by θ^j for some θ in the range (0,1). Thus if there are in all k of these backup development compounds, with u_{4j} scientists allocated during the search for the j'th compound, which therefore takes a time T_{4j}, the total expected discounted reward R under Model 1 from the $k + 1$ development compounds may be written as

$$d_1^1d_2^1d_3^1(1 + \sum_{j=1}^{k} \theta^j \prod_{i=1}^{j} d_{4i}^1)V, \tag{2}$$

where $d_i^1 = p_i exp(-\gamma_1 T_i), (i = 1, 2, 3)$, and $d_{4i}^1 = exp(-\gamma_1 T_{4i}), (i = 1, 2, ..., k)$ (d_i and d_{4i} will be used to denote the similar quantities with γ in place of γ_1).

Now writing

$$\begin{aligned} c_i &= u_i \gamma^{-1}(1 - exp(-\gamma T_i)) \qquad (i = 1, 2, 3), \\ c_{4i} &= u_{4i} \gamma^{-1}(1 - exp(-\gamma T_{4i})) \qquad (i = 1, 2, ..., k), \end{aligned}$$

and referring to equation (1), it follows that under Model 1 the total expected discounted cost C incurred in finding the $k + 1$ development compounds is

$$c_1 + d_1 c_2 + d_1 d_2 c_3 + d_1 d_2 d_3 \sum_{j=1}^{k} \prod_{i=1}^{j-1} d_{4i} c_{4j}. \tag{3}$$

For Model 2 the expressions (2) and (3) for R and C still hold if we redefine the discount factors $d_i, d_{4i}, d_i^1, d_{4i}^1$, and the costs c_i, c_{4i}, now as *expected* discount factors and *expected* costs. Dropping all subscripts, the expected discount factor for costs for a stage under Model 2 may be written

$$d = E(exp(-\gamma T)) = \int_0^\infty e(u(t)) f(x(t)) exp(-\gamma t) dt. \tag{4}$$

The expression for d^1, the expected discount factor for rewards, is of the same form with γ_1 in place of γ. Again dropping subscripts, the expected cost of a single stage under Model 2 may be written

$$\begin{aligned} c &= \int_0^\infty P(T > t) u(t) exp(-\gamma t) dt \\ &= \int_0^\infty u(t)(1 - F(x(t))) exp(-\gamma t) dt. \end{aligned} \tag{5}$$

It is also possible to describe some stages by Model 1 and others by Model 2. In this case the appropriate terms in (2) and (3) are replaced by expressions of the forms (4) and (5), with suitable suffixes.

6 Optimising the Effort Allocations

6.1 General

Since the time-scale of the development phase is similar for all projects a good approximation to a strategy which maximises profitability is to select projects for which R/C is large, and for those projects which are selected to allocate effort so as to maximise R/C, which serves as an index of profitability and will be denoted by P. Given an unlimited supply of projects for which the time-scales of research as well as development are identical this strategy would be precisely optimal. For simplicity we will from now on describe it as optimal, although strictly this is only approximately true.

¿From the expressions (2) and (3) it follows that if $\gamma_1 = \gamma$ then

$$\frac{VC}{R} = \frac{c_1}{d_1}(d_2 d_3 \sum\nolimits_R)^{-1} + \frac{c_2}{d_2}(d_3 \sum\nolimits_R)^{-1} + \frac{c_3}{d_3}\sum\nolimits_R^{-1} + \sum\nolimits_C \sum\nolimits_R^{-1}, \qquad (6)$$

where $\sum_R = 1 + \sum_{j=1}^{k} \theta^j \prod_{i=1}^{j} d_{4i}$ and $\sum_C = \sum_{j=1}^{k} \prod_{i=1}^{j-1} d_{4i} c_{4j}$. To maximise P we must minimise the right-hand side of (6). This is with respect to u_1, u_2, u_3 and u_{4i} $(i = 1, 2, ..., k)$, including the value of k. For Model 1, u_1, u_2, u_3 and u_{4i} are constants. For Model 2 they are functions of t.

Since the only quantities c and d which depend on a given u are those which correspond to the same stage (and hence have the same subscript(s)) it follows from (6) that, to maximise P, u_1 must minimise c_1/d_1. Let this minimum be A_1. Then u_2 must minimise $(A_1 + c_2)/d_2$. Let this minimum be A_2. Then u_3 must minimise $(A_2 + c_3)/d_3$. Let this minimum be A_3. Then $u_{4i}\,(i = 1, 2, ..., k)$ must minimise $(A_3 + \sum_C)\sum_R^{-1}$.

Thus the complete set of optimal allocations for the different stages may be calculated one stage at a time by means of a nested sequence of calculations if $\gamma_1 = \gamma$. When $\gamma_1 > \gamma$ we might hope to find that to optimise in turn with respect to u_1, u_2, u_3 and $u_{4i}\,(i = 1, 2, ..., k)$, and if necessary to iterate, would be a good strategy. For Model 1 this turns out to be true, and an algorithm has been written which, after some adaptation, should show whether the same is true for Model 2.

References

Andersen Consulting (1998),*Re-Inventing Drug Discovery*, Executive Briefing, www.andersen.com

Bergman, S.W. and Gittins, J.C. (1985), *Statistical Methods for Pharmaceutical Research Planning*,Marcel Dekker, New York.

Boschi, R.A.A. (1982),*Modelling exploratory research.* European Journal of Operations Research, 250–259.

Gittins, J. C. and Jones, D. M. (1974), *A Dynamic Allocation Index for New-Product Chemical Research*, Cambridge University Engineering Dept. Technical Report - Mgt.Stud/TR13 (1974).

Gittins, J. C. (1989), *Multi-Armed Bandit Allocation Indices*, Wiley.

Gittins, J. C. (1994), *Indices on thin ice*, in *Probability, Statistics and Optimisation: a Tribute to Peter Whittle*, editor F.P.Kelly, Wiley.

Gittins, J.C. (1997),*Why Crash Pharmaceutical Research*, R & D Manage ment, 79–85.

Price Waterhouse Coopers (1998), *Pharma 2005 – An Industrial Revolution int R&D*, www.pwcglobal.com/pharma/

Spilker, B. (1989),*Multinational Drug Companies – Issues in Drug Discovery and Development*,Raven Press, New York.

GittJ Gittins, J. C. and Jones, D. M. (1974), *A Dynamic Allocation Index for New-Product Chemical Research*, Cambridge University Engineering Dept. Technical Report - Mgt Stud/TR13 (1974).

Gitt1 Gittins, J. C. (1989), *Multi-Armed Bandit Allocation Indices*, Wiley.

Gitt2 Gittins, J. C. (1994),*Indices on thin ice*, in *Probability, Statistics and Optimisation: a Tribute to Peter Whittle*, editor F.P.Kelly, Wiley.

Certain Characterizations of the Uniform Distribution

Hamedani, H. and Volkmer, H.

A17004

Marquette University, USA.

Abstract. Let X1,X2,...,Xn be i.i.d. random variables with an absolutely continuous (with respect to Lebesgue measure) distribution F. Denote the corresponding order statistics by X1:n¡=X2:n¡=...¡=Xn:n. If the distribution F is uniform [0,c], then the spacing Xs:n-Xr:n and the order statistic Xs-r:n are identically distributed for 1¡=r¡s¡=n,i.e.

$$Xs : n - Xr : nandXs - r : narei.d. \tag{1}$$

The present work is mainly concerned with characterizations of the uniform distribution based on (1) for some r and s.

0. Introduction.

The problem of characterizing uniform distribution based on the identical distributions of certain spacings has been studied by many authors, in particular by Arnold, Ghosh, Huang, Shimizu, and Ahsanullah (see [**?**], [**?**], [**?**]).

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with an absolutely continuous (with respect to the Lebesgue measure) distribution function F. Denote the corresponding order statistics by $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$. If the distribution F is uniform $[0, c]$, then the spacing $X_{s:n} - X_{r:n}$ and the order statistic $X_{s-r:n}$ are identically distributed for $1 \leq r < s \leq n$, i.e.

$$X_{s:n} - X_{r:n} \sim X_{s-r:n}. \tag{0.1}$$

The present work is mainly concerned with characterizations of the uniform distribution based on (0.1) for some r and s. Sections 2 – 6 lead to improvements as well as generalizations of some results and characterization theorems of [**?**] and [**?**]. We also present a different proof of Shimizu and Huang's Theorem, [**?**], which we believe is much simpler. Finally, we mention some open problems for interested readers.

1 Preliminary Lemmas.

It is well-known [**?**, page 9] that the pdf $f_{k:n}$ of $X_{k:n}$ is given by

$$f_{k:n}(x) = \binom{n}{k} \left(\overline{F}(x)\right)^{n-k} k\, f(x) \left(F(x)\right)^{k-1}, \tag{2}$$

where $\overline{F}(x) = 1 - F(x)$, and f is the pdf corresponding to F. The pdf $f_{r,s,n}(x)$ of the spacing $X_{s:n} - X_{r:n}$, $1 \le r < s \le n$, [**?**, page 11] vanishes for $x < 0$, and for $x \ge 0$ it is given by

$$f_{r,s,n}(x) = \frac{n!}{(n-s)!(s-r-1)!r!} \quad \times \int_{-\infty}^{\infty} f(t+x) \left(\overline{F}(t+x)\right)^{n-s} [F(t+x) - F(t)]^{s-r-1} \, d(F(t))^r. \tag{3}$$

Recall that a is called a point of increase for F if $F(a+\delta) - F(a-\delta) > 0$ for all $\delta > 0$. The support of F, denoted by Supp$F$, is the set of all points of increase for F. It is a closed set and it has no isolated points (the latter holds because we assume that F is continuous.)

Lemma 1.1. If $a \in I\!R,\ b \ge 0$ and $a,\ a+b \in$ Supp F, then $b \in$ Supp $F_{r,s,n}$ for all r, s where $F_{r,s,n}$ is the distribution function of the spacing $X_{s:n} - X_{r:n}$.

Proof. We first assume additionally that

$$0 < F(a) < F(a+b) < 1. \tag{4}$$

Now, there is $\delta > 0$ (which can be chosen arbitrary small) such that, for all $x \in [b-\delta,\ b+\delta]$ and $t \in [a-\delta,\ a+\delta]$,

$$0 < F(t) < F(t+x) < 1. \tag{5}$$

Since $F(a+b+\delta) - F(a+b-\delta) > 0$, there is an $\epsilon \in (0, \delta)$ such that

$$F(t+b+\delta) - F(t+b-\delta) > 0 \quad \text{for all} \quad t \in [a-\epsilon, a+\epsilon]. \tag{6}$$

We claim that $H(b+\delta) - H(b-\delta) > 0$ where $H = F_{r,s,n}$. If not, we have

$$\int_{b-\delta}^{b+\delta} \int_{-\infty}^{\infty} f(t+x) \left(\overline{F}(t+x)\right)^{n-s} [F(t+x) - F(t)]^{s-r-1} \, f(t) \, (F(t))^{r-1} \, dt \, dx = 0.$$

By (5),

$$\int_{b-\delta}^{b+\delta} \int_{a-\epsilon}^{a+\epsilon} f(t+x) \, f(t) \, dt \, dx = 0,$$

or

$$\int_{a-\epsilon}^{a+\epsilon} [F(t+b+\delta) - F(t+b-\delta)] \, f(t) \, dt = 0.$$

In view of (6), we have, from the last equation,

$$0 = \int_{a-\epsilon}^{a+\epsilon} f(t) \, dt = F(a+\epsilon) - F(a-\epsilon).$$

This is a contradiction since a is a point of increase of F. Therefore $H(b-\delta) < H(b+\delta)$ and since δ can be arbitrary small, $b \in$ Supp$H$.

If $a, a+b \in$ Supp$F$, we can find sequences $\{a_n\},\ \{b_n\}$ in Supp$F$ converging to a and b respectively, such that for a_n, b_n, the extra assumption (4) holds. We used here the fact that Supp$F$ has no isolated point. Therefore, by the first part of the proof, $b_n \in$ Supp$H$. Since Supp$H$ is closed, we conclude that $b \in$ Supp$H$.

Lemma 1.2. For every spacing, $0 \in$ Supp$F_{r,s,n}$.

Proof. Choose $b = 0$ in Lemma 1.1.

2 Formulation of the problem.

For the special pdf

$$f(x) = \begin{cases} \frac{1}{c} & ,0 \le x \le c \\ 0 & , \text{ otherwise,} \end{cases} \tag{7}$$

with $0 < c < \infty$, we have [**?**, page 12]

$$f_{r,s,n} = f_{s-r:n} \tag{8}$$

Conjecture. Every pdf f satisfying (8) for given r and s is of the form (7) a.e.

The following lemma is given in [**?**, Lemma 1], for the special case of $s - r = 1$.

Lemma 2.1. If the pdf f satisfies (8), the support of its distribution function F is an interval $[0, c]$, where c is a positive real number or infinity.

Proof. Let f be a solution of (8). It is clear that $F(0) = 0$. It is also easy to see that the support of F agrees with the support of the distribution function $F_{k:n}$ of the order statistic $X_{k:n}$ for any k. Therefore, by (8)

$$\text{Supp}F = \text{Supp}F_{r,s,n}.$$

By Lemma 1.2, $0 \in \text{Supp}F_{r,s,n}$ and hence $0 \in \text{Supp}F$. Assume that the statement of the lemma is false. Then there are x_1, x_2 with $0 < x_1 < x_2$ such that $F(x_1) = F(x_2) < 1$. We may assume that $x_2 \in \text{Supp}F$. Since $0 \in \text{Supp}F$ and $F(0) = 0$, there is $c \in \text{Supp}F \cap (0, x_2 - x_1)$. By Lemma 1.1, we see that $x_2 - c \in \text{Supp}F_{r,s,n} = \text{Supp}F$. This is a contradiction since $x_1 < x_2 - c < x_2$ and F is constant on the interval (x_1, x_2). The statement of the lemma is now proved.

3 Solution of the problem if F is subadditive.

Theorem 3.1. Let f be a pdf satisfying (8) for given r and s, and assume that F is subadditive on its support ($\text{Supp}F = [0, c]$ by Lemma 2.1), i.e.,

$$F(x + y) \le F(x) + F(y) \qquad \text{if } x, y, x + y \in [0, c]. \tag{9}$$

Then c is finite and f is of the form (7).

Proof. By (8) and (9), we have for $x > 0$

$$f_{s-r:n}(x) = f_{r,s,n}(x) \le \frac{n!}{(n-s)!(s-r-1)!r!} \int_0^\infty f(t+x)(\overline{F}(t+x))^{n-s} \, (F(x))^{s-r-1} \, d(F(t))^r.$$

Since $F(x) > 0$ we can cancel $(F(x))^{s-r-1}$ on both sides to obtain, after simplication of the factorials,

$$f_{1:N} \le f_{r,r+1,N}, \qquad \text{where} \quad N = n - s + r + 1.$$

Since both sides of this inequality are pdf's, we conclude that

$$f_{1:N} = f_{r,r+1,N}. \tag{10}$$

Thus, we have reduced the proof of the theorem to the special case $s - r = 1$. We now solve equation (10) but we write again n in place of N. Using the substitution $s = \frac{\overline{F}(t)}{\overline{F}(x)}$ we calculate

$$
\begin{aligned}
(\overline{F}(x))^n &= \binom{n}{r} (\overline{F}(x))^n \int_0^1 (1-s)^r \, ds^{n-r} \\
&= -\binom{n}{r} \int_x^\infty [\overline{F}(x) - \overline{F}(t)]^r \, d\,(\overline{F}(t))^{n-r} \\
&= -\binom{n}{r} \int_0^\infty [F(x+t) - F(x)]^r \, d\,(\overline{F}(t+x))^{n-r}.
\end{aligned}
$$

The variable of integration in the last integral is t. If we integrate equation (8) over x from x to ∞, we obtain

$$
\begin{aligned}
(\overline{F}(x))^n = \binom{n}{r} \int_0^\infty \left(\overline{F}(t+x)\right)^{n-r} \, d(F(t))^r = \\
-\binom{n}{r} \int_0^\infty (F(t))^r \, d\, \left(\overline{F}(t+x)\right)^{n-r}.
\end{aligned} \tag{11}
$$

This equation can now be written as

$$
\int_0^\infty \left[(F(t+x) - F(x))^r - (F(t))^r\right] \, d\, \left(\overline{F}(t+x)\right)^{n-r} = 0. \tag{12}
$$

In view of (9) we have, from (12),

$$
F(t+x) = F(t) + F(x) \qquad \text{if} \quad t, x, t+x \in [0, c]. \tag{13}
$$

This implies that $F(t) = at$ for $t \in [0, c]$. Hence c is finite. Since $F(c) = 1$, we obtain $a = \frac{1}{c}$, which completes the proof.

Remark 3.2. Theorem 3.1 improves a similar result given in [**?**] in two directions: The support of F is not assumed to be finite and s need not be equal to $r + 1$.

For the special case of $r = 1,\ s = 2$, the assumption of subadditivity of F can be dropped, as shown in [**?**]. Here we provide a much simpler proof for this special case.

Theorem SH. Uniform distribution is the only absolutely continuous distribution whose $X_{2:n} - X_{1:n}$ and $X_{1:n}$ are identically distributed.

Proof. By Lemma 2.1, Supp$F = [0, c]$. Assume that Supp$F = [0, \infty]$. From (8) we have

$$
\left(\overline{F}(x)\right)^n = n \int_0^\infty \left(\overline{F}(t+x)\right)^{n-1} f(t)\, dt, \quad x \geq 0. \tag{14}
$$

This implies that

$$
\left(\overline{F}(x)\right)^n \geq n\, F(u)\, \left(\overline{F}(u+x)\right)^{n-1}, \quad x \geq 0,\ u \geq 0. \tag{15}
$$

Again by (8)

$$f(0)\ \left(\overline{F}(0)\right)^{n-1} = (n-1)\ \int_0^\infty (F(t))^{n-2}\,(f(t))^2\,dt,$$

which implies $f(0) > 0$. Since f is lower semicontinuous at 0 (this is not too hard to show), there is $\delta > 0$ and $\epsilon > 0$ such that

$$f(t) > \epsilon \quad \text{for} \quad 0 < t < \delta.$$

Then

$$\begin{aligned} f(x)\ \left(\overline{F}(x)\right)^{n-1} &= (n-1)\ \int_0^\infty \left(\overline{F}(t+x)\right)^{n-2}\ f(t+x)\,f(t)\,dt \\ &\geq (n-1)\ \int_0^\delta \left(\overline{F}(t+x)\right)^{n-2}\ f(t+x)\,f(t)\,dt \\ &\geq \epsilon\,(n-1)\ \int_0^\delta \left[\overline{F}(t+x)\right]^{n-2}\ f(t+x)\,dt \\ &= \epsilon\ \left[\left(\overline{F}(x)\right)^{n-1} - \left(\overline{F}(x+\delta)\right)^{n-1}\right]. \end{aligned}$$

Thus

$$f(x) \geq \epsilon\ \left(1 - \frac{\left(\overline{F}(x+\delta)\right)^{n-1}}{\left(\overline{F}(x)\right)^{n-1}}\right),$$

and in view of (15)

$$f(x) \geq \epsilon\ \left(1 - \frac{1}{n}\cdot\frac{\overline{F}(x)}{\overline{F}(\delta)}\right) \quad \text{for} \quad x > 0.$$

This is a contradiction since f is integrable. Thus $\mathrm{Supp}F = [0, c]$ where $c < \infty$.

4 Solution of the problem if F is superadditive.

Theorem 4.1. Left f be a pdf satisfying (8) for given r and s, and assume that F is superadditive on its support ($\mathrm{Supp}F = [0, c]$ by Lemma 2.1), i.e.,

$$F(x+y) \geq F(x) + F(y) \quad \text{if} \quad x, y,\ x+y \in [0, c]. \tag{16}$$

Then c is finite and f is of the form (7).

Proof. The proof is almost the same as that of Theorem 3.1. There is, however, a minor simplification: one can conclude that c is finite directly from (16).

Remark 4.2. The above theorem is a generalization of Theorem 1 of [**?**] to arbitrary r and s.

5 Solution of the problem if F is symmetric and $s = r + 1$.

Let $s = r + 1$. Then from (8) we have

$$n\, f(x)\, \left(\overline{F}(x)\right)^{n-1} = \binom{n}{r} (n-r) \quad \times \int_0^\infty f(t+x)\, \left(\overline{F}(t+x)\right)^{n-r-1}\, d(F(t))^r \quad \text{for} \quad x \geq 0. \tag{17}$$

Lemma 5.1. Let f be a pdf satisfying equation (8) with $s = r + 1$. Then the limit of the difference quotient $\frac{F(x)}{x}$ exists as $0 < x \to 0$ (it may be infinity).

Proof. Let f be a solution of equation (17). This equation has the form

$$n\, f(x)\, \left(\overline{F}(x)\right)^{n-1} = \int_0^\infty g(t+x)\, h(t)\, dt \tag{18}$$

with nonnegative integrable functions

$$g(t) = \binom{n}{r} (n-r)\, f(t)\, \left(\overline{F}(t)\right)^{n-r-1},$$

$$h(t) = f(t)\, r\, (F(t))^{r-1}.$$

Therefore, f is lower semicontinuous on $[0, c)$ and (18) holds for all $x \geq 0$ (not just a.e.). If $f(0) = \infty$, this gives $\lim_{x \to 0+} \frac{F(x)}{x} = \infty$. Now assume that $f(0)$ is finite. Since $\overline{F}$ is decreasing and F is increasing, we have

$$g(t+x)\, h(t) \leq \sqrt{g(t+x)\, h(t+x)\, g(t)\, h(t)}.$$

Hence

$$nf(x)\, \left(\overline{F}(x)\right)^{n-1} - nf(0) \leq \int_0^\infty \left(\sqrt{g(t+x)h(t+x)} - \sqrt{g(t)h(t)}\right) \sqrt{g(t)h(t)}\, dt. \tag{19}$$

Since $f(0)$ is finite, $\sqrt{gh}$ is square-integrable. Using the Cauchy-Schwarz inequality we find that the right-hand side of the inequality (19) tends to 0 as $x \to 0$. Since f is lower semicontinuous at 0, this implies that f is continous at 0. Therefore, $F(x) = \int_0^x f(t)\, dt$ is differentiable at 0.

Theorem 5.2. If the pdf f satisfies (8) with $s = r + 1$ and is symmetric on its support $[0, c]$ $(0 < c < \infty)$, that is, $f(c - x) = f(x)$, then f is of the form (7).

Proof. Let f be a solution of (17). Because of the symmetry, the substitution $y = c - x$ leads from (12) to

$$0 = \int_0^y \left[(F(y) - F(y-t))^r - (F(t))^r\right]\, d(F(y-t))^{n-r}. \tag{20}$$

By Lemma 5.1, $F'(0)$ exists (it may be infinity). Let $\alpha \in (0, F'(0))$. Assume that the graph of $u = F(t)$, $t \in (0, c]$, intersects the line $u = \alpha t$. Then there is $y \in (0, c]$ such that

$$F(y) = \alpha y, \quad F(t) > \alpha t \quad \text{for} \quad 0 < t < y.$$

This implies that

$$F(y) - F(t) < \alpha(y-t) < F(y-t) \quad \text{for} \quad t \in (0,y).$$

This is, however, impossible because of (20) and the fact that $\mathrm{Supp}\overline{F} = [0,c]$. Hence $F(t) \geq \alpha t$ for $t \in [0,c]$. Since α is an arbitrary number between 0 and $F'(0)$, this shows that $F'(0)$ is finite and $F(t) \geq F'(0)\,t$ for $t \in [0,c]$. Similarly, one shows that $F(t) \leq F'(0)\,t$. Hence $F(t) = F'(0)\,t$ for $t \in [0,c]$ which completes the proof.

6 Solution of the problem if F is symmetric and r = 1, s = n.

Let $r = 1$ and $s = n$. Then equation (8) can be written as

$$n(n-1)\,f(x)\,\overline{F}(x)\,(F(x))^{n-2} = n(n-1)\int_0^\infty f(t)f(t+x)\,[F(t+x)-F(t)]^{n-2}\,dt. \tag{21}$$

If we integrate over x from 0 to x we obtain

$$n\,(F(x))^{n-1}\,\overline{F}(x) + (F(x))^n = n\int_0^\infty [F(x+t)-F(t)]^{n-1}\,f(t)\,dt. \tag{22}$$

We can write this also as

$$(F(x))^{n-1}\,\overline{F}(x) = \int_0^\infty [F(t+x)-F(t)]^{n-1}\,f(t+x)\,dt \tag{23}$$

and

$$0 = \int_0^\infty \left[(F(t+x)-F(t))^{n-1} - (F(x))^{n-1}\right] f(t+x)\,dt. \tag{24}$$

Lemma 6.1. Let f be a pdf satisfying (21) and such that c is finite (see Lemma 2.1). If f is not the uniform density function, then

$$\liminf_{x\to 0^+} \frac{F(x)}{x} > \frac{1}{c}.$$

Proof. We divide both sides of (22) by nx^{n-1} and then form the limit inferior of both sides as $x \to 0^+$. Fatou's lemma gives

$$\left(\liminf_{x\to 0^+} \frac{F(x)}{x}\right)^{n-1} \geq \int_0^\infty (f(t))^n\,dt = \int_0^c (f(t))^n\,dt. \tag{25}$$

Since f is not a constant function, Hölder's inequality gives

$$1 = \int_0^c f(t)\,dt < \left(\int_0^c (f(t))^n\,dt\right)^{\frac{1}{n}} \left(\int_0^c 1^{\frac{n}{(n-1)}}\,dt\right)^{\frac{(n-1)}{n}}.$$

Hence

$$\int_0^c (f(t))^n\,dt > c^{1-n}.$$

By (25)

$$\liminf_{x\to 0^+} \frac{F(x)}{x} > \frac{1}{c}.$$

Theorem 6.2. If the pdf f satisfies (21) and is symmetric on $[0, c]$ (with c finite), then f is of the form (7).

Proof. Assume that f is not a uniform density function. By Lemma 6.1, there is $\delta > 0$ such that $F(x) > \frac{x}{c}$ for $0 < x < \delta$. By the symmetry, we also have $F(x) < \frac{x}{c}$ for
$c - \delta < x < c$. Choose the maximal number y which is less than c and such that $F(y) = \frac{y}{c}$. Then $F(y + t) < \frac{(y+t)}{c}$ for $0 < t < c - y$ and, by the symmetry, $F(t) > \frac{t}{c}$ for $0 < t < c - y$. Hence

$$[F(y + t) - F(t)]^{n-1} - [F(y)]^{n-1} < 0 \quad \text{for} \quad 0 < t < c - y.$$

This is a contradiction to equation (24) for $x = y$, which proves the theorem.

Remark 6.3. Theorem 6.2 greatly improves Theorem 2.3 of [**?**].

Theorem 6.4. Let X be a positive random variable having an absolutely continuous distribution function F. If the pdf f is strictly monotone on Supp$F$ and

$$X_{i:n} - X_{i-1:n} \sim X_{i+1:n} - X_{i:n} \quad \text{for some} \quad i, \tag{26}$$

then f is of the form (7).

Proof. By (26)

$$\begin{aligned}(i+1)\binom{n}{i+1}\int_0^\infty \left(\overline{F}(t+x)\right)^{n-(i+1)} (F(t))^i f(t)\,dt \qquad (27)\\ = i\binom{n}{i}\int_0^\infty \left(\overline{F}(t+x)\right)^{n-i} (F(t))^{i-1} f(t)\,dt,\ x>0 \quad \text{a.e.}\end{aligned}$$

Upon integration by parts on the right-hand-side of (27) we obtain

$$\int_0^\infty (F(t))^i \left(\overline{F}(t+x)\right)^{n-i-1} [f(t+x) - f(t)]\,dt = 0,\ x > 0 \quad \text{a.e.} \tag{28}$$

In view of the fact that f is strictly monotone, we have from (28)

$$f(t + x) - f(t) = 0 \quad \text{for} \quad x > 0,\ t > 0 \text{ a.e.}$$

This shows that

$$f(x) = 0 \quad \text{for} \quad x > \text{ some constant } c.$$

Otherwise

$$\int_0^\infty f(t + x)\,dt = \int_0^\infty f(t)\,dt = 1$$

or

$$F(x) = 0 \quad \text{a.e.} \quad x > 0,$$

which is obviously a contradiction. Thus f is constant on $[0, c]$.

Remark 6.5. Theorem 6.4 improves Theorem 2.2 of [**?**].

7 Summary.

We have solved our problem if F is either subadditive or superadditive for any $1 \leq r < s \leq n$. Under the assumption of symmetry we have solved the problem only for the cases $s = r+1$ and $r = 1,\ s = n$. It would be worth-while to investigate the symmetric case for arbitrary r and s.

M. Ahsanullah, *On characterizations of the uniform distribution based on functions of order statistics, Aligrah J. Stat.* **9** (1989), 1–6.
H. David, *Order Statistics*, Wiley & Sons, New York, 1981.
J. S. Huang, B. Arnold and M. Ghosh, *On characterizations of the uniform distribution based on identically distributed spacings,* Sankhya, **41** (1979), 109–115.
R. Shimizu and J. S. Huang, *On a characteristic property of the uniform distribution, Ann. Inst. Statist. Math.* **35** (1983), Part A, 91–94.

Nonparametric covariance analysis in field experiments

Jose, C. T.

A17015

Central Plantation Crops Research Institute, India.

Abstract. In this paper, a method is proposed to estimate/eliminate positional effect in field experiments nonparametrically. A semiparametric regression model with treatment effect as the parametric component and the positional/location effect (covariate) as a bivariate nonparametric function has been used to analyse the field experimental data. The only assumption about the positional effect is that it is a smooth spatial (bivariate) function. The method is also extended to analyse the data in the presence of treatment x position interaction effect. The proposed method is illustrated through a simulation study.

Keywords. Covariance Analysis, Kernel Smoothers, Local Linear Regression, Nonparametric Regression, Semiparametric Regression.

1 Introduction

Nonparametric modeling technique is a rapidly growing and exciting branch of statistics in recent years because of the recent theoretical developments and the widespread use of the fast and inexpensive computers. In this paper we discuss its applications in field experiments. We generally use block designs in field experiments to control the experimental error due to positional variations. The underlying assumption in classical block designs that the homogeneity of experimental area within the block may not satisfy always, particularly when the block size is large. Also we may not know in advance the soil fertility gradient and other factors influencing the response variable to divide the experimental area into homogeneous blocks. The treatment x block interaction effect is usually taken as experimental error in the analysis of block designs and wherever this interaction effect is present, the experimental error will be high. In the present study, semiparametric technique has been used to estimate/eliminate the positional effect. The treatment effect is the parametric component and the positional/location effect is the covariate, which is taken as a bivariate nonparametric function. The only assumption about the positional effect is that it is a smooth spatial (bivariate) function. The method is also extended to analyse the data in the presence of treatment x position interaction effect. The proposed method is illustrated through a simulation study.

2 Model Settings and estimators

A semiparametic regression model with treatment effects as parametric component and the positional effect (covariate) as nonparametric component is considered for the field experiment. The semiparametric model is given by

$$Y = X\beta + f(U, V) + \varepsilon$$

(1) Where $Y = [Y_1, Y_2, ..., Y_n]^T$ is the observation vector which is taken as the deviation from the mean,$X = [X_1, X_2, ..., X_n]^T$ is the design matrix, $\beta = [\beta_1, \beta_2, ..., \beta_p]^T$ is the treatment effect vector,$f(U, V) = [f(U_1, V_1), f(U_2, V_2), ..., f(U_n, V_n)]$ is the nonparametric spatial function representing the positional effect and ε is the iid random error vector with mean zero. It is assumed that $f(U, V)$ is a smooth function and $\sum f(U_i, V_i) = 0$. Backfiting algorithm is used to estimate the treatment and positional effect in the regression model and the estimates are given by

$$\beta = (X^T(I - S)X)^{-1}X^T(I - S)Y \quad and \quad f = S(Y - X\beta)$$

Where, S is the smoothing matrix derived using local linear regression (Ruppert and Wand, 1994). Let SUV be the row of the smoother matrix correspond to the smoother vector S_{UV}^T evaluated at the observation point $(U, V) = (U_1, V_1), (U_2, V_2), ..., (U_n, V_n)$. Then,

$$S[S_{U_1V_1}...S_{U_nV_n}]^T$$

where,

$$S_{uv}^T = e_1^T(Z_{uv}^T W_{uv} Zuv)^{-1} Z_{uv}^T W_{uv}$$

with, $e_1^T = [100]$,$W_{uv} = diag\{k[(\frac{U_1-u}{h_1}), (\frac{V_1-v}{h_2})], ..., k[(\frac{U_n-u}{h_1}), (\frac{V_n-v}{h_2})]\}$ for some bivariate kernel functions K and bandwidths h_1 and h_2 and

$$Z_{uk} = \begin{bmatrix} 1 & (u_1 - u) & (V_1 - v) \\ \vdots & \vdots & \vdots \\ 1 & (U_n - u) & (V_n - v) \end{bmatrix}$$

Under the assumption that the treatments are allotted at random to the spatial locations, it can be shown thatβ is asymptotically unbiased and its asymptotic variance is $\sigma^2(X^TX)^{-1}$ which is same as when the model is fully parametric. An estimate of σ^2 is given by

$$\sigma^2 = \frac{1}{(n - k - 1 - trace(S))}[y - X\beta - f]^f[y - X\beta - f]$$

The variance of is obtained by

$$V(\beta) = PP^T\sigma^2$$

where,$P = (X^T(I - S)X)^{-1}X^T(I - S)$. Under some regularity conditions, it can be proved that β is$\sqrt{n}$ -consistent (0psomer and Ruppert, 1999). The significance of the positional effect f can be tested using the lack-of-fit test (Hart, 1997). In model (1), it is assumed that the treatment x position interaction effect is absent and in the presence of this effect the model (1) can be modified as follows

$$y_{ij} = f_i(U_{ij}, V_{ij}) + \varepsilon_{ij} \quad , \quad i = 1, 2, ..., k; \quad j = 1, 2, ...ni; \quad \sum n_i = n$$

Where,y_{ij} is the observed value of the of the i^{th} treatment at the spatial location (U_{ij}, V_{ij}), $f_i(U_{ij}, V_{ij})$ is the expected value of the i^{th} treatment at the spatial location(U_{ij}, V_{ij}) and ε_{ij} is the iid random error with mean zero. We assume that the mean functions$f_i(.), i = 1, , k$ are smooth. Let $Y^* = [y_{11}, y_{12}, ..., y_{1n_1}y_{21}, y_{22}, ..., y_{kn_k}]^T$ be the rearranged observation vector,$Y_i = [y_{i1}, ..., y_{in_i}]^T$, andF_i is the n_ix1 vector

$[f_i(U_{i1}, V_{i1}), ..., fi(U_{ini}, V_{ini})]^T$. Then the model (2) can be written in matrix form as

$$Y^* = \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix} = \begin{bmatrix} F_1 \\ \vdots \\ F_k \end{bmatrix} + \varepsilon^*$$

or

$$Y^* = F + \varepsilon^*$$

(3) Where, $\varepsilon^*$is the error component corresponding to Y^*. The solution to the above regression problem can be obtained as$F = S^*Y^*$ Where, the smoothing matrix S^* is given by

$$S^* = \begin{bmatrix} S_1 & 0 & \cdots & 0 \\ 0 & S_2 & \cdots & 0 \\ \vdots & & & \\ & & \cdots & \\ 0 & 0 & \cdots & S_k \end{bmatrix}$$

and S_i is the $n_i x n_i$ smoother matrix for the observations (U_{ij}, V_{ij}), $j = 1, 2, ...n_i$. An estimate of σ^2 is given by

$$\sigma^2 = \frac{1}{(n - k - 1 - trace(S^*))}[F^* - F]^T[Y^* - F]$$

The variance of is obtained by

$$V(F) = S^* S^{*T} \sigma^2$$

The significance of the treatment x position interaction effect can be tested by comparing the fitted models of the equations (1) and (2) using the lack of fit test (Hart, 1997). In many situations, the number of experimental units may comparatively small and estimating the spatial function using the bivariate smoother will be inadequate. In such situations, bivariate additive model can be fitted instead of the two dimensional spatial function used in models (1) and (2). By using bivariate additive function, the model (1) can be modified as

$$Y = X\beta + f_1(U) + f_2(V) + \varepsilon \qquad (3)$$

Where, f1 and f2 are the univariate nonparametric function representing the effect of the U and V directions and it is assumed that $\sum f_1(U_i) = \sum f_2(V_i) = 0$. Let$M_1$ andM_2 are the centered smoother matrices corresponding to U and V. The backfitting algorithm will provide an explicit solution to the above semiparametric regression model and the estimates are given by $\beta = (X^T(I-Q)X)^{-1}X^T(I-Q)Y$ and $f = f_1 + f_2 = Q(Y - X\beta)$ The estimates f_1 ,f_2 and Q are obtained by solving the set of equations

$$\begin{bmatrix} I & M_1 \\ M_2 & I \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} (Y - X\beta)$$

$$f_1 = \{I - (I - M_1M_2)^{-1}(I - M_1)\}(Y - X\beta) = Q_1(Y - X\beta)$$

$$f_2 = \{I - (I - M_2M_1)^{-1}(I - M_2)\}(Y - X\beta) = Q_2(Y - X\beta)$$

and

$$Q = Q_1 + Q_2$$

An estimate of ? 2 is given by

$$\sigma^2 = \frac{1}{(n-k-1-trace(Q))}[y - X\beta - f]^T[y - X\beta - f]$$

Ignoring the bias, an approximate ?-level pointwise confidence band around the estimated function f is given by $f(U_i, V_i) = z_{\alpha/2}\sigma\sqrt{[QQ^T]_{ii}}\ for i = 1, ..., n$, where, $[QQ^T]_{ii}$ represents the element in the ii^{th} position of the matrix $[QQ^T]$

3 Simulation Study

A simulation study is carried out to see the practical implications of the theoretical results given in Section 2. For the simulation study, we considered the following model

$$y_{ij} = f_i(u_{ij}, v_{ij}) + \varepsilon_{ij} \qquad i = 1, k; j = 1, ..ni; \sum ni = n$$

wherey_{ij} is the j^{th} observation of the ith treatment,(u_{ij}, v_{ij}) is the spatial location and ? is the mean zero random error. In this study we have taken k=4, n=400,$f_1(u, v) = 2$, $f_2(u, v) = f_3(u, v) = 2(2 + \sin(u + v))andf_4(u, v) = 2(3 + \sin(3u + 2v))$. The spatial locations of the 400 observations are obtained by dividing the region [0,1]x[0,1] equally andε is taken as $N(0, \sigma^2)$. The treatments are allotted randomly to the spatial locations. Based on the above 100 sets of data were simulated for different values ofσ and the functions f_i's and the error variance ? were estimated using the method given in Section 2. The Mean Squared Errors (MSE) of the estimated values with the true values of 100 sets of simulated data were worked out (Table 1). It can be observed that the estimates were very close to the true values. The MSE varies with change in error variance and the choice of bandwidths.

The optimum bandwidth (bandwidth corresponds to the minimum MSE) will depend on the error variance and the curvature of the function. It can be observed from the table that the optimum bandwidth is more when the error variance is large. The optimum bandwidth is comparatively small for the functions with large curvature (f4) than the functions with small curvature (f1, f2 or f3). The optimum bandwidth can be obtained using the method of cross-validation (Hardle, 1990).

Conclusion

We generally use block designs to eliminate positional effect in field experiments. The underlying assumption of homogeneity with in the block may not be true in many situations. Also in classical block designs, we are taking block x treatment interaction effect as experimental error. Whenever the above assumptions fail, the experimental error become very large. In the present study, a method is proposed to eliminate the positional effect nonparametrically and the only assumption about the positional effect is that it is a smooth spatial (bivariate) function. The method is also extended to analyse the data in the presence of treatment x position interaction effect.

Bandwidth		MSE of the estimates multiplide by 1000				
σ	$(h_1 = h_2)$	σ	f_1	f_2	f_3	$f4$
0.25	0.20	0.12	11.70	15.44	16.31	17.15
	0.25	0.16	10.93	11.77	10.61	13.79
	0.30	0.12	5.95	6.27	6.95	15.58
	0.35	0.18	7.72	36.65	6.21	24.58
0.50	0.20	0.58	58.5	55.31	54.67	55.86
	0.25	0.43	36.33	37.10	37.27	37.09
	0.30	0.34	31.81	27.20	25.28	36.06
	0.35	0.26	24.17	21.55	24.36	42.47
0.75	0.20	0.60	116.96	112.21	110.18	122.18
	0.25	0.55	65.23	69.92.	65.18	72.77
	0.30	0.48	62.05	61.47	62.00	70.94
	0.35	0.43	49.65	55.91	52.74	61.11
1.00	0.20	2.01	219.70	212.14	218.3248.44	
	0.25	2.11	149.29	141.68	141.24	174.16
	0.30	1.67	116.30	115.46	111.49	150.46
	0.35	1.39	83.33	86.09	86.86	96.32

References

Hardle, W. (1990).*Applied Nonparametric Regression*. Cambridge Universit Press.

Hart, J.D. (1997).*Nonparametric smoothing and lack-of-fit tests*. Springer Verlag, New York.

Opsomer and Ruppert, D. (1999).*A Root-n Consistent Estimator for Semiparametric Additive Modeling*, Journal of Computational and Graphical Statistics, 8, 715- 732.

Ruppert, D. and Wand, M.P. (1994).*Multivariate Locally Weighted Least Squares Regression*. Annals of Statistics, 22, 1346-70.

Breast Cancer Recurrences

Khoshbin, E.[1] **and Davies, R.**[2]

A11025

[1] Department of Mathematics, Statistics and Computer Science, Tehran University, Iran.
[2] Centre for Applied Statistics, Fylde College, Lancaster University, UK.

Abstract. In this paper we consider durations from an epidemiological study of breast cancer. The first duration for each patient is the time from initial treatment to recurrence of the disease. For many women this duration will be right-censored. The second duration is the time to second recurrence for those women who responded to treatment after first recurrence.

1 Introduction

Worldwide, over half a million women develop breast cancer each year but half of all these cases occur in North America and Europe, which contain less than one-fifth of the female world population . There are several types of breast cancer, some slow-growing, some aggressive. Those that go undetected or untreated can spread to surrounding breast tissue, then to the lymph nodes under the arm, and then to other parts of the body in a process known as metastasis.

In England and Wales, approximately 25000 women will develop carcinoma of the breast annually and at the Christie Hospital in the city of Manchester, UK, where the data used in this paper were collected , about 1600 women are registered with a diagnosis of breast carcinoma each year (Dos Santos 1994).

A number of factors have been identified which increase a woman's chance of developing breast cancer: increasing age, late child-bearing (first child after the age of 30 years), nulliparity (no children), early menarche, late menopause , family history (first degree relative, e.g. sister or mother, particularly premenopausal), obesity, and ionizing radiation.

Four types of treatment (surgery, radiotherapy, hormone therapy, and chemotherapy) are used in the management of breast cancer. The women will, in general, receive an adjuvant therapy or therapies (radiotherapy and/or adjuvant treatment, for instance, Tamoxifen) to surgery following initial diagnosis. Three different types of surgery are used in the management of breast cancer: minor surgery, simple mastectomies, and radical mastectomies.

Section 2 introduces the dataset providing information about the covariates and some simple statistics about the data. It also discusses the substantive context of the study. Initial data analysis is reported in section 3. Model formulation and the model fitting results are presented and discussed in section 4. Section 5 conclude.

2 The Data

The data used in this research cover the women referred to the Christie Hospital, UK, with breast cancer between 1980 and 1985, and their subsequent monitoring

until July 1991. This dataset was used by Dos Santos (1994) and Dos Santos et. al. (1995), in a study of recurrence following treatment for breast cancer. The youngest patient in the study was aged 21 years, the oldest 88, their mean age was 56.6 years (standard deviation 13.1 years). Out of the 917 women in the sample, more than half (513) had no recurrence by the end of the study. The two durations of interest are the durations, measured in years, from primary treatment to first recurrence (Duration 1), and from response to treatment following first recurrence to second recurrence (Duration 2). No recurrence, death from other causes and loss from the study were treated as right-censoring.

The explanatory variables used in the analysis are AGE, STAGE of tumour at diagnosis with three levels measuring the severity of the disease, and treatment (TREAT) with three levels : treatment1-minor surgery; treatment2 -simple mastectomy; and treatment 3 - radical mastectomy. It is emphasized that treatment always includes one of these three surgical interventions.

The classification that was used in order to assign a patient clinically to a stage that would be easy to monitor subsequently was based on the largest diameter of the lump that can be felt by clinical examination. The classification is used consistently as follows:

Stage 1 - Tumour size $\leq$ 2 cm and no nodes involved;

Stage 2 - Tumour size $>$ 2 cm but $\leq$ 5 cm and no nodes or tumour size $\leq$ 5 cm with

nodes;

Stage 3 - Tumour size $>$ 5 cm with or without nodes.

After dropping those patients with bilateral disease (left and right breast cancer) and those whose measurements for stage were missing (27 of them), the sample size is 917 women and this is the total number of patients considered in the analysis.

2.1 Initial Data Analysis

Piecewise exponential plots of the log-hazards are shown in Figures (1) and (2) for the first and second recurrence respectively. These were produced using GLIM macro Phaz (Francis et al., 1993), with no explanatory variables. Although the piecewise log-hazards are widely dispersed, there is some evidence in both plots of a "sickle" -shape with the hazard first increasing and then decreasing. This is a common feature of hazard rates in medical research and is consistent with the result of (Dos Santos et al. 1994), in their analysis of these data. Kaplan Meier plots for survival functions for first and second recurrences are shown in Figure 3; these plots confirm the poor outcomes for second recurrences. The complementry log-log plots for the Kaplan Meier survival function are shown in Figure 4. Both are non-linear providing further evidence of a non-monotonic hzard rate. First and second recurrence duration are plotted against each other in Figure 5. The wide scatter of points provides no visual evidence of any systematic relationship.

To give clearer evidence about the shape of the hazard function when controlling for covariates, a number of standard "two parameter" survival models were fitted to both the first and second durations using GLIM. The model fitting results are shown in Tables 2 and 3. The best fits are provided by the log-normal for both

durations, confirming that the hazards are sickle shaped. The log-logistic is the second best in each case. For duration 2, the log-logistic shape parameter is greater than 1 indicating a sickle shape. However, for duration 1, the log-logistic indicates a monotonically declining hazard although the fit is substantially worse than for the log-normal. The Weibull and Gompertz distributions with their monotonic hazards provide decidedly worse fits.

The structural parameter estimates are generally as expected. In particular, the hazard increases (for the Weibull and Gompertz) and the expected duration to recurrence (for the log-logistic and the log-normal) correspondingly decreases with the stage of the disease at initial diagnosis; the smaller the tumour, the better the prognosis. Moreover, recurrence is less likely for older women.

The estimated treatment effects present a more complex pattern. Recurrence appears to be more likely for treatment 2 (simple mastectomy) than for the reference treatment category (minor surgery). In interpreting this result, it is important to note that treatment has not been randomized and it is plausible that minor surgery tends to be used for the least threatening cases. More reassuringly, treatment 3 (radical mastectomy) reduces the hazard (or correspondingly increases the expected duration) for duration 1. However, it is estimated to have a disadvantageous effect on second duration. We would speculate that this is a "sample selection" effect whereby those patients with the extreme surgery treatment of radical mastectomy and recurrence tend to have particularly pernicious cancers.

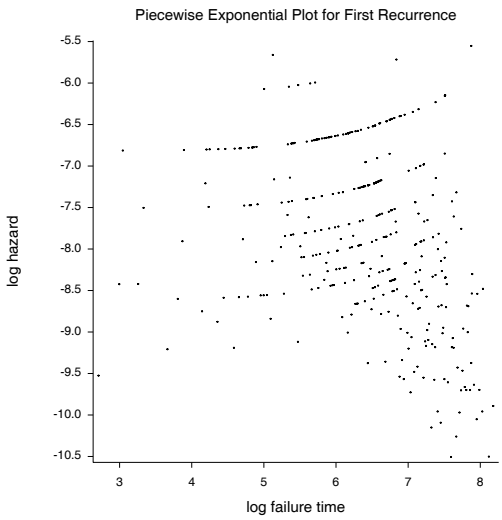

Fig. 1. Piecewise exponential plot for log-hazard to first recurrence (time to failure measured in days). Both scales use natural logarithms.

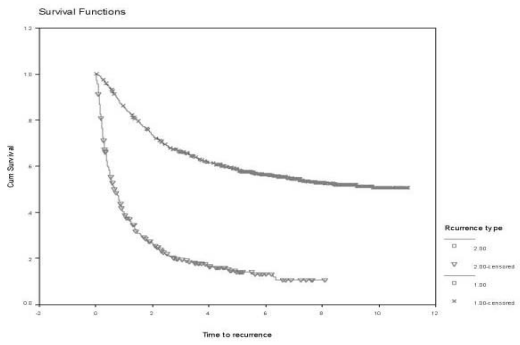

Fig. 2. Kaplan Meier graph for both recurrence (time to recurrence measured in years).

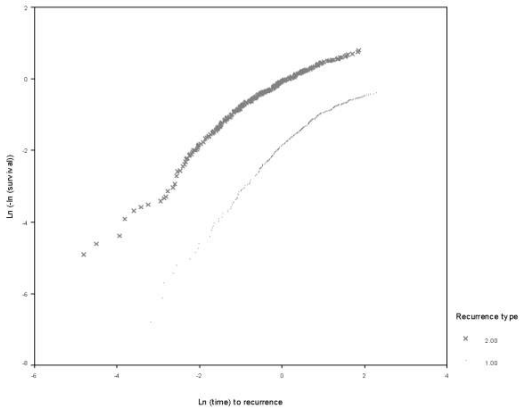

Fig. 3. Complementary log-log transformation of Kaplan Meier estimates (time to recurrence measured in years).

3 Statistical Modelling

3.1 Model Formulation

The initial data analysis of the previous section indicates that the marginal distributions of the t_1 and t_2 have sickle-shaped hazards, but we have no substantive theory to inform us of the likely shapes of the conditional distributions. For this kind of data, it is possible that the cancer is never eliminated and presents an increasing risk of recurrence with time at an individual level. The sickle-shaped hazard at an aggregate level would then be explained by the sample selection effects of hetero-

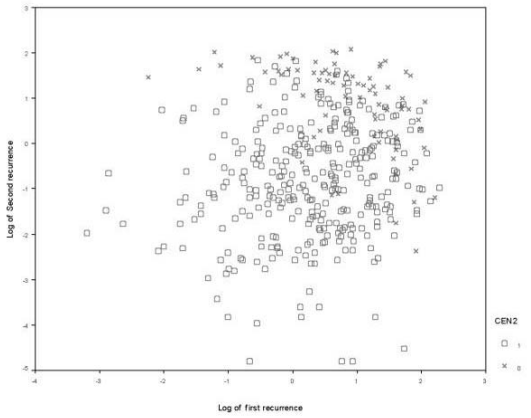

Fig. 4. First and second recurrence durations are plotted against each other (time to recurrence measured in years). Both scales use natural logarithms.

	Weibull Model(PH)				Log-Logistic Model(AL)			
	Duration 1		Duration 2		Duration 1		Duration 2	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
shape	0.786	.034	.721	.031	.984	.043	1.106	0.049
constant	-2.037	.246	-.405	.273	2.027	.322	-0.229	.412
age*10^{-2}	-.629	.377	-.853	.446	.901	.519	1.302	.649
Stage2	0.785	.109	.422	.123	-1.163	.152	-0.596	.174
Stage3	1.567	.184	.663	.204	-2.046	.282	-0.792	.282
Treatment2	0.305	.122	.473	.154	-0.379	.160	-0.702	.206
Treatment3	-0.259	.153	.675	.177	0.354	.195	-0.792	.246
LOG-LIKE	-1303.20		-426.75		-1284.50		-403.84	

Table 1. Model fitting results for conventional Weibull and log-logistic models

	Log Normal Model(AL)				Gompertz Model(PH)			
	Duration 1		Duration 2		Duration 1		Duration 2	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Shape	1.745	.068	1.564	0.634	-.228	0.025	-.522	.059
Constant	2.009	.324	-0.223	.406	-1.768	.246	.056	.281
Age*10^{-2}	.975	.515	1.316	.646	-.678	.378	-.915	.449
Stage2	-1.133	.152	-.555	.176	.770	.109	.428	.122
Stage3	-2.119	.295	-.711	.298	1.492	.184	.618	.203
Treatment2	-.356	.159	-.721	.205	.315	.122	.470	.153
Treatment3	.355	.194	-.806	.249	-.155	.153	.652	.177
LOG-LIKE	-1273.20		-403.20		-1638.50		-679.07	

Table 2. Model fitting results for conventional Log-normal and Gompertz models.

geneity; with time, those with higher levels of frailty will suffer recurrence leaving the least susceptible in remission. On the other hand, it is possible that a regenerative process does begin to protect against recurrence after a period, the hazard accordingly having a sickle-shape at both the individual and aggregate levels.

To deal with these two possible situations , we use two different families of models. To represent sickle-shape hazards for the conditional distributions, we propose log-logistic formulations. These are preferred over the alternative log-normal because (i) they are easier to handle analytically, (ii) they permit a monotonically declining hazard which is not entirely ruled out as a possibility for duration 1 , and (iii) they include very similar shapes to the log-normal. To represent monotonically increasing hazards we use the conventional Weibull distribution. We note that the Weibull also permits a monotonically declining hazard.

3.2 Log-logistic Log-logistic Normal Model (LLN Model)

In this subsection we specify our LLN model, with a log-logistic distribution for the first and second durations and a Normal distribution for the random effect. With this specification we can write the densities for the two durations as follows

$$f_1\left(t_{i1}\left|\mathbf{x_i}, \upsilon_{\,i}\right.\right) = \frac{\alpha_1}{\theta_{i1}} \frac{\left(t_{i1}/\theta_{i1}\right)^{\alpha_1 - 1}}{\left[1 + \left(t_{i1}/\theta_{i1}\right)^{\alpha_1}\right]^2}$$

$$f_2\left(t_{i2}\left|t_{i1}, \mathbf{x_i}, \upsilon_{\mathbf{i}}\right.\right) = \frac{\alpha_2}{\theta_{i2}} \frac{\left(t_{i2}/\theta_{i2}\right)^{\alpha_2 - 1}}{\left[1 + \left(t_{i2}/\theta_{i2}\right)^{\alpha_2}\right]^2}$$

where $\theta_{i1} = \exp\left(\sum_{r=1}^{R} \beta_{1r} x_{i1r} + \varepsilon_{\mathbf{i}}\right)$ and $\theta_{i2} = \exp\left(\sum_{r=1}^{R} \beta_{2r} x_{i2r} + t_1 + \varepsilon_i\right), \upsilon_i = exp\left(\varepsilon_i\right)$ represent the random effect for ith individual, $\alpha_1$and α_2 are shape parameters. The likelihood for this model is

$$L = \prod_{i=1}^{n} \int f\left(t_{i1}, t_{i2}\left|\mathbf{x_i}\right., \upsilon_i\right) dF\left(\upsilon_i\right)$$

$$= \prod_{i=1}^{n} \int f_1\left(t_{i1}\left|\mathbf{x_i}\right., \upsilon_i\right) f_2\left(t_{i2}\left|\mathbf{x_i}, \mathbf{t_{i1}}, \upsilon_{\mathbf{i}}\right.\right) dF\left(\upsilon_i\right).$$

Since there is no close form for this integral we have to use the quadrature method again by changing the integral to a summation over quadrature points. When there is a failure in both durations, the contribution of the i-th likelihood is given by

$$L_i = \sum_{k=1}^{m} \left\{ \frac{\alpha_1}{\theta_{i1}} \frac{\left(t_{i1}/\theta_{i1}\right)^{\alpha_1 - 1}}{\left[1 + \left(t_{i1}/\theta_{i1}\right)^{\alpha_1}\right]^2} \right\} \left\{ \frac{\alpha_2}{\theta_{i2}} \frac{\left(t_{i2}/\theta_{i2}\right)^{\alpha_2 - 1}}{\left[1 + \left(t_{i2}/\theta_{i2}\right)^{\alpha_2}\right]^2} \right\} P_k,$$

where the (ξ_k, P_k) are the quadrature locations and masses for the numerical integration (although we also use a closed form approach), and m is the number of

quadrature points, where for $m = 4, 6, 8$, the log-likelihood turns out to be the same.

When there is no failure in the second duration, the contribution of the i-th likelihood is

$$L_i = \sum_{k=1}^{m} \left\{ \frac{\alpha_1}{\theta_{i1}} \frac{(t_{i1}/\theta_{i1})^{\alpha_1 - 1}}{\left[1 + (t_{i1}/\theta_{i1})^{\alpha_1}\right]^2} \right\} \frac{1}{1 + (t_{i2}/\theta_{i2})^{\alpha_2}} \; P_k.$$

If treatment is not successful following first recurrence , the contribution to the likelihood is given by

$$L_i = \sum_{k=1}^{m} \frac{\alpha_1}{\theta_{i1}} \frac{(t_{i1}/\theta_{i1})^{\alpha_1 - 1}}{\left[1 + (t_{i1}/\theta_{i1})^{\alpha_1}\right]^2} P_k.$$

The last situation occurs when there is no failure in the first duration. The contribution of the i-th likelihood for this situation is

$$L_i = \sum_{k=1}^{m} \; \frac{1}{1 + (t_{i1}/\theta_{i1})^{\alpha_1}} P_k.$$

3.3 Weibull Weibull Gamma (Burr Distribution)

In this subsection we specify the WWG model, with Weibull distributions for the first and second durations and a gamma distribution for the random effect. The gamma distribution is a common model for a multiplicative frailty term υ (Clayton 1978; Oakes 1982 ; Lindley and Singpurvalla 1986; Clayton and Cuzik 1985). It has the density

$$g(\upsilon) = \frac{\upsilon^{k-1} \exp(-\upsilon)}{\Gamma(k)} k > 0.$$

If the mean of the frailty distribution is specified as k. With a constant term in the linear predictor, this does not result in any loss in generality. Specifically, writing

$$\lambda = \exp\left(\beta' \mathbf{x} + \varepsilon\right),$$

where $\upsilon = \exp(\varepsilon)$ is the frailty random effect, we may alternatively and equivalently write

$$\lambda = \upsilon \exp\left(\beta' \mathbf{x}\right).$$

The advantage of the gamma is that it is conjugate with the Weibull, giving the Burr distribution. This conjugate property is not confined to univariate duration distributions but extends to a range of multivariate distributions, including the bivariate Weibull considered in this subsection. The main advantage is computational; numerical integration is not required in fitting the model. In this application,

we therefore follow the common practice in medical statistic of assuming a gamma frailty distribution.

We specify the Weibull location parameters as

$$\lambda_{i1}^{*} = \exp\left(\beta_{10} + \sum_{r=1}^{R} \beta_{1r}\mathbf{x_{ir}} + \ \varepsilon_{\mathbf{i}}\right) = \upsilon_i \exp\left(\beta_{10} + \sum_{r=1}^{R} \beta_{1r}\mathbf{x_{ir}}\right) = \upsilon_i \lambda_{i1},$$

and

$$\lambda_{i2}^{*} = \exp\left(\beta_{20} + \sum_{r=1}^{R} \beta_{2r}\mathbf{x_{ir}} + \mathbf{t_{t1}} + \ \varepsilon_{\mathbf{i}}\right) = \upsilon_i \exp\left(\beta_{20} + \sum_{r=1}^{R} \beta_{2r} x_{ir} + t_{i1}\right) = \upsilon_i \lambda_{i2},$$

where R is the number of covariates. The model is given by

$$f\left(t_{i1}, t_{i2} \left|\mathbf{x_i}\right.\right) = \int f_1\left(t_{i1} \left|\mathbf{x_i}, \upsilon_{\mathbf{i}}\right.\right) f_2\left(t_{i2} \left|t_{i1}, \mathbf{x_i}, \upsilon_{\mathbf{i}}\right.\right) dG\left(\upsilon_i\right)$$

and the likelihood is

$$\begin{aligned} L_i = & \int \upsilon_i \lambda_{i1} \gamma_1 t_{i1}^{\gamma_1 - 1} \exp\left(-\upsilon_i \lambda_{i1} t_{i1}^{\gamma_1}\right) \\ & \times \upsilon_i \lambda_{i2} \gamma_2 t_{i2}^{\gamma_2 - 1} \exp\left(-\upsilon_i \lambda_{i2} t_{i2}^{\gamma_2}\right) \frac{\upsilon_i^{k-1} \exp\left(-\upsilon_i\right)}{\Gamma\left(k\right)} d\upsilon_i \\ = & \left(k+1\right)\left(k\right)\left[\lambda_{i1}\gamma_1 t_{i1}^{\gamma_1 - 1} \lambda_{i2}\gamma_2 t_{i2}^{\gamma_2 - 1}\right] \\ & \times \left(1 + \lambda_{i1} t_{i1}^{\gamma_1} + \lambda_{i2} t_{i2}^{\gamma_2}\right)^{-k-2} \end{aligned}$$

If the second recurrence time is right-censored , the contribution to the likelihood is given by

$$\begin{aligned} L_i = & \int f_1\left(t_{i1} \mid \mathbf{x_i}, \theta_{\mathbf{i}}\right) S_2\left(t_{i2} \mid t_{i1}, \mathbf{x_i}, \theta_{\mathbf{i}}\right) dG\left(\theta_i\right) \\ & = \int \theta_i \lambda_{i1} \gamma_1 t_{i1}^{\gamma_1 - 1} \exp\left(-\theta_i \lambda_{i1} t_{i1}^{\gamma_1}\right) \\ & \times \exp\left(-\theta_i \lambda_{i2} t_{i2}^{\gamma_2}\right) \frac{\theta_i^{k-1} \exp\left(-\theta_i\right)}{\Gamma\left(k\right)} \\ & = k \lambda_{i1} \gamma_1 t_{i1}^{\gamma_1 - 1} \left(1 + \lambda_{i1} t_{i1}^{\gamma_1} + \lambda_{i2} t_{i2}^{\gamma_2}\right)^{-k-1} \end{aligned}$$

If treatment is not successful following first recurrence , the contribution to the likelihood is given by

$$\begin{aligned} L_i = & \int f_1\left(t_{i1} \mid \mathbf{x_i}, \theta_{\mathbf{i}}\right) dG\left(\theta_i\right) d \\ & = \int \theta_i \lambda_{i1} \gamma_1 t_{i1}^{\gamma_1 - 1} \exp\left(-\theta_i \lambda_{i1} t_{i1}^{\gamma_1}\right) \frac{\theta_i^{k-1} \exp\left(-\theta_i\right)}{\Gamma\left(k\right)} \\ & = k \lambda_{i1} \gamma_1 t_{i1}^{\gamma_1 - 1} \left(1 + \lambda_{i1} t_{i1}^{\gamma_1}\right)^{-k-1} \end{aligned}$$

Finally, if the first duration is right-censored, the contribution to the likelihood is given by

$$L_i = \int S_1\left(t_{i1} \mid \mathbf{x_i}, \upsilon_{\mathbf{i}}\right) dG\left(\upsilon_i\right)$$

$$\begin{aligned} L_i &= \int \exp\left(-\upsilon_i \lambda_{i1} t_{i1}^{\gamma_1}\right) \frac{\upsilon_i^{k-1} \exp\left(-\upsilon_i\right)}{\Gamma\left(k\right)} d\upsilon_i \\ &= \left(1 + \lambda_{i1} t_{i1}^{\gamma_1}\right)^{-k}. \end{aligned}$$

3.4 Model Fitting Results

The model fitting results for the full LLN (log-logistic ,log-logistic,Normal) and WWG (Weibull Weibull Gamma) models are shown in Table 4. NAG algorithm (1993) E04UCF was used to maximize the log-likelihood function in each case. The parameter estimates from the homogeneous models reported in section 2 were used as starting values for the modified Newton-Ralphson algorithm. It is emphasized that, following the conventional parametrization, the explanatory variables scale the hazard in the WWG model but scale the mean in the LLN model. This explains the different signs of most of the parameter estimates.

With a difference in the log-likelihoods of only 0.30, the WWG model provides a very marginally better fit to the data than the LLN model. The WWG model indicates an increasing hazard at an individual level for survival times to first and second recurrences although, for second recurrence, the estimated shape parameter is only marginally greater than 1 (the value for a constant hazard) suggesting a low rate of increase. The slightly worse fitting LLN model gives a sickle-shape hazard for both durations. On the basis of these two similarly fitting models, we therefore conclude whether the conditional distributions for the times to first and second recurrences have increasing, sickle or even differently shaped, hazards.

The Weibull and log-logistic conditional hazards for time to first recurrence are plotted in Figure **??** with age set to 40 years, stage 2 tumour size, and treatment 2. The same plots for time to second recurrence are shown in Figure **??** with duration 1 assumed to be 1.5 years. Both plots show alarmingly different patterns for the Weibull and log-logistic models.

The marginal hazards for first and second recurrence are compared for the Burr model in Figure **??**. The explanatory variables were given the same values as for Figures **??** and **??**. These plots are consistent with the earlier evidence of sickle-shapes hazards and the higher hazards for second recurrence. However, they also suggest that after about two years the hazards have converged.

The marginally better fitting model (WWG:Burr) is compared with various simplified versions in Table 5. This table also includes p-values for the parameter estimates for the full model. These were calculated from likelihood ratio test statistics and required refitting the model with each covariate removed, in turn. This method was adopted because of difficulty experienced in recovering the Hessian matrix from the NAG(1993) routine E04UCF (see Khoshbin). The p-values indicate that STAGE is highly significant and that the age effects are not significant at least for duration to first recurrence. Treatment 2 is marginally significant (at the

10% level) and treatment 3 is not significantly different from treatment 1 for duration to first recurrence, but treatment 2 is significantly different from treatment 1 for duration to second recurrence. Treatment 3 is significantly different at the 10% level.

The model denoted by WWG-t_1 is the full model without the Markov-type effect. The large reduction in the log-likelihood (corresponding to LR $X^2 = 10.78$ with 1 degree of freedom) provides strong evidence for the dependence of t_2 on t_1 after controlling for frailty. The direction of the effect is unexpected: longer durations to first recurrence are associated with higher hazards for second recurrence. Moreover , we are unable to explain the increase in the random effect parameter on dropping the Markov-type variable. The observed increase would be expected if the Markov variable represented a positive dependence, as is usually the case; some of the positive dependence would be picked up by the frailty terms which would therefore have an increased variance. We would not have expected this to happen when the Markov dependence is negative. However, we note that there are complex inter-relationships between the two durations for these data which may make it difficult to find simple explanations for some of the model sensitivities. The marginal (Burr) hazards for this model are plotted in figure 9. The general shapes are similar to those for the full model with Markov effect and, in particular, the locations of the maxima are approximately the same.

The model denoted by WW+t_1 is a homogeneous model (i.e. no frailty) with Markov dependence. The reduction in the log-likelihood of circa 55 in comparison to the full model provides unequivocal evidence of frailty effects even though a likelihood ratio test is not strictly appropriate because the simplified model lies at the boundary of the parameter space of the full model. With two exceptions, the estimated values of the coefficients of the covariates are substantially lower in absolute value than the corresponding estimates for the full model. This attenuation of parameter estimates when frailty is omitted is consistent with the theoretical results of Lancaster and Nickell (1980).

One exception is the coefficient for the treatment 3 covariate (radical mastectomy) for the second duration . The anomalous result for this covariate in a homogeneous model has already been discussed in section 2. It was argued that, since treatments have not been randomized, it is possible that minor surgery tends to be used for the least threatening cases and treatment 3 for the worst. The results for the full model are more plausible and tend to confirm that at least part of the explanation lies in the failure to allow for frailty; with frailty explicitly modelled in the full model, the treatment 3 effect is not significantly different from that for treatment 1 at the 5% level and is estimated to have a lower hazard than treatment 2. The full model results are still surprising in estimating lower hazards for treatment 1 (minor surgery) than for treatment 2 (simple mastectomy). But this pattern is consistent across duration 1 and duration 2 and may be due to treatment 1 being used primarily for cases with a good prognosis.

The second exception to the attenuation of parameter estimates in the WW+t_1 model is the coefficient for the Markov-type variable t_1 ; it is negative while the corresponding estimate for the full model is positive. However, this marked discrepancy is consistent with theoretical expectations for Markov-type variables. As noted, for example, by Massey et al. (1970), ignoring frailty will tend to induce a spurious positive Markov dependence. In effect, the Markov variable acts as a "proxy" vari-

Duration 1	**LLN+t_1**	**WWG+t_1**
gamma	1.512	1.498
constant	2.142	-0.466
Age$\times 10^{-2}$	0.779	-1.062
Stage2	-1.131	1.471
Stage3	-1.999	2.401
Treat2	-0.413	0.535
Treat3	0.311	-0.416
Duration 2		
gamma	1.391	1.066
constant	0.922	0.465
Age$\times 10^{-2}$	1.675	- 2.097
Stage2	-1.032	1.298
Stage3	-1.528	2.136
Treat2	-0.932	1.040
Treat3	-0. 740	0.729
t_1	-0.089	0.241
Random effect	1.261	0.247
LOG-LIKE	-1672.31	-1672.01

Table 3. Results for the Log-Logistic/ Log-Logistic/ Normal (LLN+t1), and Burr (WWG+t1)models. (With only one scale parameter).

able for the temporal dependence due to the omitted frailty. As already noted, the Markov variable in the full model indicates a negative dependence, with duration to first recurrence therefore having a positive effect on the hazard (and hence a negative effect on duration) of second recurrence. It appears that the spurious positive Markov dependence created by ignoring frailty in the WW+t_1 model exceeds this "true" negative dependence and thereby results in a potentially misleading net positive dependence (indicated by the negative estimated effect on the hazard).

4 Conclusion

Substantively, we cannot draw firm conclusions from the analyses in this paper because of problems over the robustness of the models investigated. Nevertheless, some tentative conclusions are in order.

First, there appear to be distinctively different processes governing the durations to first and second recurrences. The hazards are different, especially at short durations; the chances of a rapid recurrence are much higher for second recurrence. The explanatory variables also have different impacts. For example, the stage of the disease at diagnosis appears to have rather less impact on the second recurrence.

Second, not only do the marginal hazards have a sickle shape, but there is some evidence that the conditional hazards may also be sickle shaped in that the log-logistic formulations perform better overall than the Weibull formulations in the random effect models. This could be an important result. If the conditional hazards are Weibull with a monotonically increasing hazard, then each patient is heading

Duration 1	WW	WW+t_1	WWG-t_1	WWG+t_1	*p value*
Variables	homogeneous	Markov	Frailty	Full model	
gamma	0.790	.790	1.255	1.498	
constant	-2.055	-2.054	- 1.047	-.466	
age$\times 10^{-2}$	-0.6	-.6	-0. 863	- 1.062	0. 32
stage2	0.775	0.775	1. 343	1.471	2.781e-06
stage3	1.566	1.566	2. 194	2. 401	3.292e-05
treat2	0.309	0.309	0. 499	0. 535	0.097
treat3	-.256	-.256	-0. 376	-. 416	0.313
Duration 2					
gamma	0.723	0.721	1. 042	1.066	
constant	-0.406	-.117	.512	0.465	
age$\times 10^{-2}$	-0.009	-0.010	-1.915	-2.097	0.066
stage2	0.414	0.341	.981	1.298	7.089e-06
stage3	0.663	0.598	1.701	2.136	2.044e-04
treat2	0.481	0.490	.946	1. 040	3.245e-03
treat3	0.685	0.781	.889	0. 729	0. 094
dur1	-	-0.104	-	0.241	0.020
random effect	-		.371	0. 247	1.499e-13
LOG-LIKE	-1730.96	-1726.58	-1 677.40	- 1672.01	

Table 4. Results for nested Weibull models. tww: WW- Homogeneous Weibull model for each duration WW+t1- As WW with first duration as a covariate for second duration WWG-t1- Heterogeneous (Gamma) extension of WW (Bivariate Burr Model) WWG+t1- As WWG-t1 with first duration as a covariate for second duration.

for a recurrence sooner or later, although, of course, she could die through old age before experiencing recurrence if she has a low "frailty" measure. On the other hand, if hazards are sickle-shaped this suggests that there is a tissue regeneration process with at least some of the patients moving towards a negligible chance of recurrence in due course.

Methodologically, the breast cancer analyses have revealed the expected type of relationship between the individual level (conditional) hazards and the aggregate (marginal) hazards, within a more complex modelling context than that required in the previous work (Khoshbin). They have demonstrated also the attenuation of parameter estimates to be expected on theoretical grounds when fitting models without random effects.

The other methodological lessons to be learned from this paper are less precise and concern the robustness of random effect models. Our results tend to confirm that sensitivity to parametric specification in random effect models can be an indication of more serious misspecifications. Extended models attempting to disentangle these various effects proved to be highly sensitive to the parameterization adopted.

Duration 1	**LLN-t_1**	**LLN+t_1**	**WWN+t_1**	**WWN-t_1**
gamma	1.702	1. 703	.891	0.824
constant	2.121	2.121	-2. 290	-2.115
Age$\times 10^{-2}$	0.787	0.788	-.692	-0.709
Stage2	-1.082	-1.081	.919	0.828
Stage3	-2.082	-2.082	1. 730	1.610
Treat2	-0.417	-0.417	0.305	0.312
Treat3	0.311	0.310	-0.327	-0.285
Randorm effect$_1$	1.379	1.380	-0.734	-.436
Duration 2				
gamma	1.192	1.190	1. 523	1.578
constant	0.381	0.365	1.734	-.163
Age$\times 10^{-2}$	1.506	1.504	-2.236	-0.731
Stage2	-0.794	-0.788	0.425	0.820
Stage	-1.285	-1.277	.323	0.941
Treat2	-0.844	-.842	1.070	0.892
Treat3	-0.817	-0.818	1.981	1.532
t_1	-	.003	-0.379	-
Random effect$_2$	0.762	.751	2.202	2.151
LOG-LIKE	-1671.46	-1671.45	- 1692.36	-1709.33

Table 5. Model fitting results with two scale parameters for the Log-Logistic/Log-Logistic/Normal, and Weibull/Weibull/Normal, with and without Markov type effect..

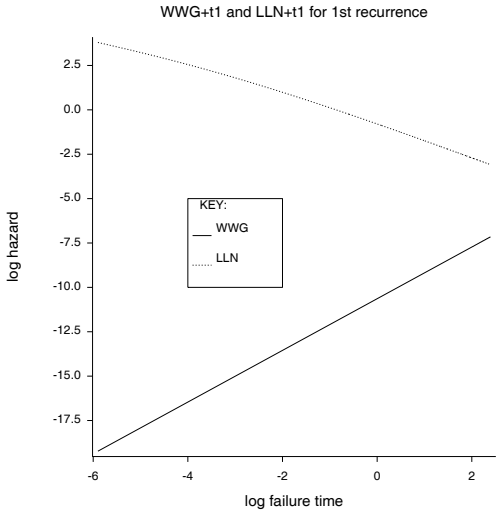

Fig. 5. Plot log hazard versus log failure time for first recurrence. The solid line(——) is Burr specification and dotted line(.....) is log-logistic/log-logistic/normal, (time to failure measured in years). Both scales use natural logarithms.

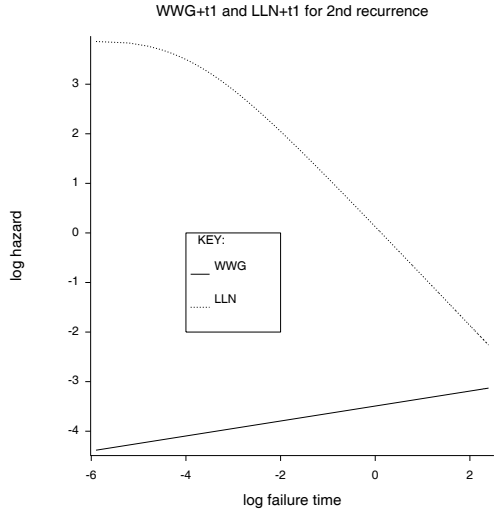

Fig. 6. Plot of log hazard of duration 2 versus log failure time with Burr specification with Markov effect ,solid line(—), and then plot log-logistic/log-logistic/normal specification with Markov effect, dotted line(.....), (time to failure measured in years). Both scales use natural logarithms.

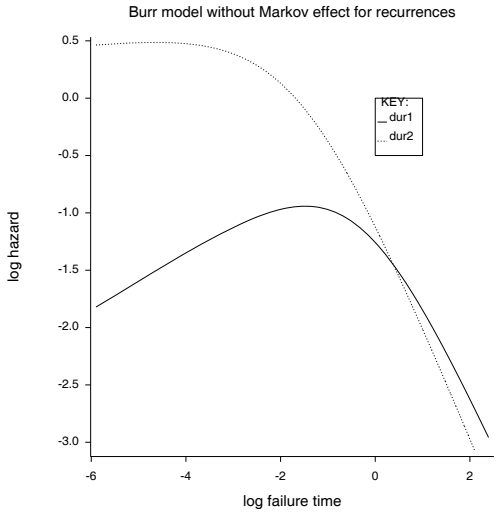

Fig. 7. Log hazard of duration 1 and duration 2 versus log failure time with Burr specification (Model 3 of chapter 3), (time to failure measured in years). Both scales use natural logarithms.

References

Clayton, D. (1978). A Model For Association In Bivariate Life-Tables. Biometrika **65**, 141-151.

Clayton, D. & Cusick, J. (1985). Multivariate generalizations of the proportional hazard model(with Discussion). J. R. Statist. Soc. **A 148** 82-117.

Dos Santos, D, (1994). Distinguishing the Effects of Time Dependence from the Effects of Frailty in Multiple Spell Breast Cancer Data. Thesis ,Centre for Applied Statistics, Lancaster University.

Dos Santos, D, Davies. R. B., and Francis, B. (1995). Nonparametric hazard versus nonparametric frailty distribution in modelling recurrence of breast cancer. J*ournal of Statistical Planning and Inference* **47**, 111-127.

Francis, B. Green, M. Payne, C. (1993). Glim4. The Statistical System for Generalized Linear Interactive Modelling. Clarendon Press. Oxford.

Khoshbin, E. Modelling Two Stage Duration Process: Thesis, Centre for Applied Statistics, Lancaster University.

Khoshbin, E. Davies R.B. Mitchell J. D. Relationship between age of onset and speed of progression of MND. Working paper (2001).

Lancaster, T. and Nickell, S. (1980). The analysis of Re-Employment Probabilities for the Unemployed. *Journal of the Royal Statistical Society.* A **134**: 141-165.

Lindley, D. and Singpurvalla N. (1986). Multivariate distributions for the life length. *Journal of Appled Prob.*, **23**, 418-431.

Massy, W. F., Montgomery, D. B. and Morrison, D. G. (1970). Stochastic models of buying behaviour. MIT Press, (Cambridge, Mass).

NAG (1993) Numerical Algorithms Group library manual, mark 16 (Oxford, England).

Oakes, D. (1982). A model for association in bivariate data. Journal of the Royal Statistical Soc., B 44, 414-422.

Improving on the MLE of a Mean of a Spherical Distribution

Marchand, E.

P17011

University of New Brunswick, Canada.

Abstract. For the problem of estimating under squared error loss the mean of a p-variate spherically symmetric distribution where the mean lies in a ball of radius m, a sufficient condition for an estimator to dominate the maximum likelihood estimator is obtained. We use this condition to show that the Bayes estimator with respect to a uniform prior on the boundary of the parameter space dominates the maximum likelihood estimator whenever $m \leq \sqrt{p}$ in the case of a multivariate student distribution with d degrees of freedom, $d \geq p$. The sufficient condition $m \leq \sqrt{p}$ matches the one obtained by Marchand and Perron (2001) in the normal case with identity matrix. Furthermore, we derive a class of estimators which, for $m < \sqrt{p}$, dominates the maximum likelihood estimator simultaneously for the normal distribution with identity matrix and for all multivariate student distributions with d degrees of freedom, $d \geq p$. The family of distributions where dominance occurs includes the normal case; and includes all student distributions with d degrees of freedom, $d \geq 1$, for the case $p = 1$.

Keywords. Maximum Likelihood Estimator, Restricted Parameter Space, Squared Error Loss, Dominance, Simultaneous Dominance, Spherically Symmetric Distribution, Scale Mixture of Normals, Multivariate Student Distribution.

1 Introduction

Consider the problem of estimating under squared error loss the mean θ of a spherically symmetric distribution, based on the observation X and with the constrained parameter space $\Theta(m) = \{\theta \in R^p : \|\theta\| \leq m\}$ for some m fixed, $m > 0$. In the normal case with identity covariance matrix, Marchand and Perron (2001) showed that the Bayes estimator δ_{BU} with respect to the boundary uniform prior on $\partial\Theta(m)$ dominates the maximum likelihood estimator δ_{mle} whenever $m \leq \sqrt{p}$. An interesting question is whether a similar result holds for other spherically symmetric distributions. This is indeed the objective of our research and, moreover, we focus on the multivariate student distribution which represents perhaps one of the most important alternatives to the normal model and permits us, through its scale mixture of normals representation, to give explicit results.

The starting point in our inquiry is a sufficient condition (Theorem 1) for an estimator to dominate δ_{mle}, which was implicitly given by Marchand and Perron (2001, Theorem 3), and which is applicable in general to spherically symmetric distributions. We then study how this condition applies to δ_{BU} and obtain further specifications for the multivariate student case with d degrees of freedom. We establish in Section 3 (Example 1) that the condition $m \leq \sqrt{p}$ is, whenever $d \geq p$, once again sufficient for δ_{BU} to dominate δ_{mle}. The common sufficient condition is interesting and somewhat surprising, in view of its simplicity, and the fact that both the functional form of the estimator δ_{BU} and the distribution under which the risks are evaluated vary with d.

We also can view the sufficient condition for dominance of Theorem 1 as a sufficient condition for simultaneous dominance (Theorem 2), meaning a condition under which a single estimator δ_0 dominates δ_{mle} simultaneously for a subfamily of spherical distributions. Of course, it is the hope that such a simultaneous condition of dominance can be made explicit for important subfamilies of spherical distributions, possibly including the normal case. Simultaneous dominance is an appealing property in view of the intrinsic motivation of assessing or searching for procedures that retain good or optimal properties over a range of probability models. Although there seems to be a relative paucity of results in this direction, this is not a new theme; for instance, some recent work on estimating a multivariate mean (without constraints) has dealt with procedures that perform well not only for the normal model, but also for a range of spherical or elliptical models. As an example, Cellier and Fourdrinier (1995), gave a class of estimators that dominate the unbiased estimator, for $p \geq 3$, simultaneously for all spherically symmetric distributions subject to (weak) risk finiteness conditions.

In the second part of Section 3, we focus again on multivariate student distributions with d degrees of freedom and obtain two examples of simultaneous dominance. In particular, we obtain an explicit estimator δ_0 which, for $m < \sqrt{p}$, dominates δ_{mle} simultaneously for all multivariate student distributions with $d \geq p$ as well as the normal distribution with identity covariance matrix. This is a particular interesting result since no theoretical elements that we know of guaranteed the existence of such a simultaneously dominating δ_0. The simultaneous dominating estimators obtained, although simple, may well fail to be attractive for a given single distribution, but the result permits us to envisage locally (or globally) more attractive estimators to enjoy the same simultaneous dominating property.

Before proceeding in Section 3 with these dominance results, we pursue with by collecting some further notations, definitions and properties for later use.

2 DEFINITIONS AND PRELIMINARIES

Throughout, we shall denote $\|x\|$ and $\|\theta\|$ by r and λ respectively. We consider distributions with probability density functions

$$f_\theta(x) = h(\|x - \theta\|) \tag{1}$$

where h is such that $h(t) < h(0)$ for all $t > 0$. For such distributions, the maximum likelihood estimator of θ is uniquely given by $\delta_{\text{mle}}(x) = \left(\frac{m}{\|x\|} \wedge 1\right) x$. The function $g_{h,\lambda}(r) = \mathrm{E}_\theta[\frac{\theta' X}{\|X\|} | \|X\| = r]$ plays a pivotal role in our dominance results as it intervenes in both (i) the decomposition of risks (see Theorem 1), and (ii) the functional form of the Bayes estimator $\delta_{g_{h,\lambda}}$ with respect to a uniform prior on the sphere $\{\theta : \|\theta\| = \lambda\}$, given by (e.g., Marchand, 1993, proof of Theorem 2.3.)

$$\delta_{g_{h,\lambda}}(x) = \frac{1}{r} g_{h,\lambda}(r) x \ .$$

Of particular interest is the boundary uniform prior, and the associated Bayes estimator $\delta_{BU} = \delta_{g_{h,m}}$ which was shown by Marchand and Perron (2001) to dominate δ_{mle} in the normal case with identity covariance matrix whenever $m \leq \sqrt{p}$.

We further define $\bar{g}_{h,m}(r) = \sup_{0\leq\lambda\leq m} g_{h,\lambda}(r)$, and $A_{h,m} = \{r > 0 : \bar{g}_{h,m}(r) < r\}$.

Remark. ¿From its definition and the Cauchy-Schwarz inequality, it is easy to see that $g_{h,\lambda}(r) < \lambda \leq m$. Hence $A_{h,m}$ always contains the set $[m, \infty)$.

Our conditions for dominance in Section 3 below are given first implicitly in terms of $\bar{g}_{h,m}$ and $A_{h,m}$ (Theorems 1 and 2), and we proceed by developing more explicit conditions in the multivariate student case (Theorems 3 and 4). In order to achieve this, we require two technical lemmas. We begin with an expression for $g_{h,\lambda}$ for scale mixture of normals where X admits the representation:

$$\mathcal{L}(X|V = v) = \mathrm{N}_p(\theta, v^{-1}I_p)\ ; \tag{2}$$

for some positive random variable V.

Lemma Marchand, 1993 *For scale mixture of normals as defined above, we have*

$$g_{h,\lambda}(r) = \lambda \frac{E[I_{\frac{p}{2}}(tV)e^{-sV}V]}{E[I_{\frac{p}{2}-1}(tV)e^{-sV}V]}\ ;$$

where $t = \lambda r$, $s = (\lambda^2 + r^2)/2$, and $I_\nu(y)$; $\nu \geq -1/2, y \geq 0$; is the modified Bessel function of order ν given by $I_\nu(y) = \sum_{i\geq 0} \frac{(\frac{y}{2})^{\nu+2i}}{i!\,\Gamma(i+\nu+1)}$.

Remark. When referring to a specific distribution of the mixing parameter in (2) for which $\mathrm{E}[V] < \infty$, we can assume without loss of generality (and we will hereafter) that $\mathrm{E}[V] = 1$. This is so since, whenever $\mathrm{E}[V] \neq 1$ and $\mathrm{E}[V] < \infty$, we can always transform the problem to work with: (i) the observation $X^* = \sqrt{\mathrm{E}[V]}X$, (ii) the constraint $\|\mathrm{E}[X^*]\| \leq m^*$ with $m^* = m\sqrt{\mathrm{E}[V]}$, and (iii) the representation for X^* as in (2) (with mean $\theta^* = \theta\sqrt{\mathrm{E}[V]}$) corresponding to the mixing parameter $V^* =^d \frac{V}{E[V]}$, for which $\mathrm{E}[V^*] = 1$.

We now continue with a further representation of $g_{h,\lambda}(r)$ for the cases in (2) where $\mathcal{L}(V) = \mathrm{Gamma}(a, b)$. As mentioned in Remark 2, there is no loss of generality in limiting ourselves to the cases where $\mathrm{E}[V] = 1$; i.e., $a = b$; which corresponds to the multivariate student cases with degrees of freedom $d, d > 0$, with $a = b = \frac{d}{2}$.

Lemma 1. *If the distribution of X follows a multivariate student distribution with d degrees of freedom and $m \leq \sqrt{d}$ then*

$$\bar{g}_{h,m}(r) = g_{h,m}(r) \tag{3}$$

$$\leq \frac{m^2 r}{m^2 + r^2 + d}[1 + (1 \vee \frac{d}{p})] \tag{4}$$

for all $r > 0$.

Proof. See the Appendix.

3 DOMINANCE RESULTS

We begin this section with a sufficient condition for an estimator $\delta_g(x) = \frac{1}{r}g(r)x$ to dominate $\delta_{\rm mle}$. The proof is essentially the same as the one given by Marchand and Perron (2001) in the normal case, but given here for sake of completeness.

3.1 General dominance results

Theorem 2. *For distributions as in (1) and $\delta_g(x) = \frac{1}{r}g(r)x$, the estimator δ_g dominates δ_{mle} as long as*

$$2\bar{g}_{h,m}(r) - (r \wedge m) < g(r) < (r \wedge m)$$

for all $r \in A_{h,m}$ and $g(r) = r$ otherwise.

Proof. We have

$$\begin{aligned} \mathrm{R}(\theta, \delta_g) &= \mathrm{E}_\theta[\|g(\|X\|)\frac{X}{\|X\|} - \theta\|^2] \\ &= \mathrm{E}_\theta[\|\theta\|^2 + g^2(\|X\|) - 2g(\|X\|)\frac{\theta' X}{\|X\|}] \\ &= \|\theta\|^2 + \mathrm{E}_\theta[\{g(\|X\|) - g_{h,\lambda}(\|X\|)\}^2 - g^2_{h,\lambda}(\|X\|)\,]. \end{aligned}$$

Hence $\mathrm{R}(\theta, \delta_{\rm mle}) - \mathrm{R}(\theta, \delta_g) = \mathrm{E}_\theta[\{g_{\rm mle}(\|X\|) - g(\|X\|)\}\{g_{\rm mle}(\|X\|) + g(\|X\|) - 2g_{h,\lambda}(\|X\|)\}]$, which is indeed positive for all $\theta \in \Theta(m)$ under the stated conditions.

□

Remark. Note that $\delta_{\bar{g}_{h,m}}$ satisfies the conditions of Theorem 1 whenever $A_{h,m} = (0, \infty)$, while its truncated version (i.e., with $g(r) = \bar{g}_{h,m}(r) \wedge g_{\rm mle}(r)$) always dominates $\delta_{\rm mle}$. In the normal case with identity covariance matrix (i.e., Marchand and Perron, 2001), it was established that (i) $\delta_{\bar{g}_{h,m}} = \delta_{BU}$; and that (ii) $A_{h,m} = (0, \infty)$ if and only if $m \leq \sqrt{p}$. Now, by requiring an estimator δ_g to fulfill the conditions of Theorem 1 for all h in a family $\mathcal{H}$ of distributions, we obtain the following simultaneous dominance result. General implications of Theorem 2 are discussed at the beginning of Section 3.3.

Theorem 3. *Let $\delta_g(x) = \frac{1}{r}g(r)x$. The estimator δ_g dominates δ_{mle} simultaneously for all $h \in \mathcal{H}$ as long as*

$$2 \sup_{h \in \mathcal{H}} \bar{g}_{h,m}(r) - (r \wedge m) < g(r) < (r \wedge m)$$

on the set $A_{\mathcal{H},m}$; and $g(r) = r$ otherwise; with $A_{\mathcal{H},m} = \{r : 2\sup_{h\in\mathcal{H}} \bar{g}_{h,m}(r) - (r \wedge m) < (r \wedge m)\}$.

3.2 Dominance results for the multivariate student distribution

To pursue, let us recall (i.e., Remark 3) the following conditions for dominance:

(A) the estimator $\delta_{\bar{g}_{h,m}}$ will dominate $\delta_{\rm mle}$ whenever $A_{h,m} = (0, \infty)$;

(B) δ_{BU} will dominate δ_{mle} whenever $A_{h,m} = (0, \infty)$ and $\bar{g}_{h,m} = g_{h,m}$.

In this section, Theorem 3 gives simple conditions for which (A) and (B) are satisfied in the multivariate student cases.

Theorem 4. *Assume that the distribution of X follows a multivariate student distribution with d degrees of freedom and $\delta_g(x) = \frac{1}{r} g(r)x$.*

a) If $m \leq \sqrt{p \wedge d}$, $d < \{1 + 2[(p/m^2 - 1) + 2\sqrt{(p/m^2 - 1)^2 + p/m^2}]\}p$ and g is such that

$$2\frac{m^2 r}{m^2 + r^2 + d}[1 + (1 \vee \frac{d}{p})] - (r \wedge m) < g(r) < (r \wedge m)$$

for all $r > 0$ then δ_g dominates δ_{mle}.

b) If $m \leq \sqrt{p}$, $d \geq \{1 + 2[(p/m^2 - 1) + 2\sqrt{(p/m^2 - 1)^2 + p/m^2}]\}p$ and g is such that $g(r) = m$ if $\{2r - m(1 + d/p)\}^2 \leq m^2(1 + d/p)^2 - 4(m^2 + d)$ and

$$2\frac{m^2 r}{m^2 + r^2 + d}[1 + \frac{d}{p}] - (r \wedge m) < g(r) < (r \wedge m)$$

otherwise then δ_g dominates δ_{mle}.

c) If $m \leq \sqrt{d}$ and g is such that

$$2g_{h,m}(r) - (r \wedge m) < g(r) < (r \wedge m)$$

for all $r \in A_{h,m}$ and $g(r) = r$ otherwise then δ_g dominates δ_{mle}.

d) If $m \leq \sqrt{p \wedge d}$ and g is such that

$$2g_{h,m}(r) - (r \wedge m) < g(r) < (r \wedge m)$$

for all $r > 0$ then δ_g dominates δ_{mle}.

Proof.

a) Here is an application of Theorem 2 and Lemma 2. We need only to verify that $\frac{m^2 r}{m^2+r^2+d}[1 + (1 \vee \frac{d}{p})] < (r \wedge m)$ for all $r > 0$. It is easy to see that $\frac{m^2 r}{m^2+r^2+d}[1 + (1 \vee \frac{d}{p})] < r$ for all $r > 0$ if and only if $m \leq \sqrt{p \wedge d}$ and $\frac{m^2 r}{m^2+r^2+d}[1+(1 \vee \frac{d}{p})] < m$ for all $r > 0$ if and only if $d < \{1 + 2[(p/m^2 - 1) + 2\sqrt{(p/m^2 - 1)^2 + p/m^2}]\}p$.

b) This proof is similar to the one of part a) except that here $\frac{m^2 r}{m^2+r^2+d}[1 + \frac{d}{p}] \geq m$ whenever $\{2r - m(1 + d/p)\}^2 \leq m^2(1 + d/p)^2 - 4(m^2 + d)$.

c) This is a direct application of Theorem 2 and Lemma 2.

d) This is similar to the case b). We need only to verify that $A_{h,m} = (0, \infty)$. From Lemma 2 we know that $\bar{g}_{h,m}(r) \leq \frac{m^2 r}{m^2+r^2+d}[1 + (1 \vee \frac{d}{p})]$ and from the proof of part a) we know that $\frac{m^2 r}{m^2+r^2+d}[1 + (1 \vee \frac{d}{p})] < r$ for all $r > 0$ if $m \leq \sqrt{p \wedge d}$ therefore $\bar{g}_{h,m}(r) < r$ for all $r > 0$ and $A_{h,m} = (0, \infty)$.

□

Example 1. The estimator δ_{BU} dominates δ_{mle} whenever $m \leq \sqrt{p \wedge d}$. In fact, this is a special case of Theorem 3, part d).

Remark. For $d \geq p$, δ_{BU} dominates δ_{mle} whenever $m \leq \sqrt{p}$, duplicating Marchand and Perron's (2001) sufficient condition in the normal case. Finally, it also can be shown that for $d \geq p$, the condition $m \leq \sqrt{p}$ is also necessary for $A_{h,m}$ to equal $(0, \infty)$. This is established by considering the necessary condition $\lim_{r \to 0} \frac{\bar{g}_{h,m}(r)}{r} \leq 1$, and using the expression (8) in the proof of Lemma 2 (see the Appendix) to infer that $\lim_{r \to 0} \frac{\bar{g}_{h,m}(r)}{r} = \frac{m^2}{m^2+d} \frac{d+p}{p}$.

3.3 Simultaneous dominance results for the multivariate student distribution

We now turn to applications of Theorem 2, that is the specification of estimators that dominate δ_{mle} for several distributions simultaneously, and results are given herein for multivariate student distributions. Note that the choice $g(r) = \sup_{h \in \mathcal{H}} \bar{g}_{h,m}(r)$ for $r \in A_{\mathcal{H},m}$ satisfies the conditions of Theorem 2, while the above results imply that $A_{\mathcal{H},m} = (0, \infty)$ for $m \leq \sqrt{p}$ with $\mathcal{H}$ being the multivariate student family with degrees of freedom $d \geq p$. However, this estimator is not given explicitly. The results below pertaining to the family of multivariate student distributions witd d degrees of freedom, $d \geq p$ are of particular interest since: (i) dominance is shown to hold as well for the normal distribution with identity covariance matrix, and (ii) the family includes all univariate student distributions with d degrees of freedoms, $d \geq 1$, whenever $p = 1$.

Theorem 5. *Assume that the distribution of X follows a multivariate student distribution with d degrees of freedom and $\delta_g(x) = \frac{1}{r} g(r) x$.*

a) If $m \leq \sqrt{d_0}$, $d_0 \leq p$ and g is such that

$$4 \frac{m^2 r}{m^2 + r^2 + d_0} - (r \wedge m) < g(r) < (r \wedge m)$$

for all $r > 0$ then δ_g dominates δ_{mle} for all d, $d_0 \leq d \leq p$.

b) If $m < \sqrt{p}$, $d_0 \geq p$ and g is such that

$$2 \frac{m^2 r}{p} (1 \vee \frac{p + d_0}{m^2 + r^2 + d_0}) - r \wedge m < g(r) < r \wedge m$$

for all $0 < r < p/m$, and $g(r) = m$ otherwise then δ_g dominates δ_{mle} for all d, $d \geq d_0$ and for the normal distribution case as well.

Proof. The proof is an application of Theorem 3.

a) Since $mr^2/(m^2 + r^2 + d)$ is a decreasing expression in d, our result satisfies the conditions of Theorem 3, part a), for all d, $d_0 \leq d \leq p$.

b) We use the results of Theorem 3, part a) and part b). Since $\frac{m^2 r}{p}(1 \vee \frac{p+d_0)}{m^2+r^2+d_0}) < r \wedge m$ for all $0 < r < p/m$ and $\frac{m^2 r[1+1 \vee d/p]}{m^2+r^2+d} \leq \frac{m^2 r}{p}(1 \vee \frac{p+d_0)}{m^2+r^2+d_0})$ for all $r > 0$, $d \geq d_0$ we obtain our result. Finally, δ_g and δ_{mle} are both bounded and the

densities converge to the one of a normal distribution as $d \to \infty$ which implies that the risk functions converge and our result is still valid for the normal distribution.

□

Example 2. Translating directly the conditions of Theorem 4b to the univariate case with $d_0 = 1$, we obtain for $m < 1$ that the estimator

$$\delta_0(x) = m\{m(1 \vee \frac{2}{m^2 + x^2 + 1}) \wedge \frac{1}{|x|}\}x$$

dominates δ_{mle} simultaneously for all student distributions with degrees of freedom $d \geq 1$ and the normal distribution as well.

APPENDIX

In the appendix we shall prove that the expressions (3) and (4) of Lemma 2 are valid whenever X follows a multivariate student distribution with d degrees of freedom and $m \leq \sqrt{d}$.

Proof: expression (3) in Lemma 2. Let $T = \theta' X/\lambda R$ so $g_{h,\lambda}(r) = \lambda \mathrm{E}_\lambda[T|R^2 = r^2]$. If we can show that $\mathrm{E}_\lambda[T|R^2 = r^2]$ is nondecreasing in λ for all r, $r > 0$ then we shall have $\bar{g}_{h,m} = g_{h,m}$. Moreover, the expression $\mathrm{E}_\lambda[T|R^2 = r^2]$ will be nondecreasing in λ for all r, $r > 0$ if the conditional distribution of T given that $R^2 = r^2$ has monotone likelihood ratio in T for all $r > 0$, where λ is the parameter and r is fixed. Let $f_\theta(x) = h(\|x - \theta\|)$ as in expression (1).

If $p = 1$ then T is a discrete random variable taking the values -1 and 1, R^2 is an continuous random variable on $(0, \infty)$ and their likelihood is given by $\varphi_{\lambda,1}$ with

$$\varphi_{\lambda,1}(t, r^2) = \frac{1}{2r} h(\sqrt{r^2 + \lambda^2 - 2\lambda r t}).$$

Similarly, if $p > 1$ then T, R^2 have a joint density $\varphi_{\lambda,p}$ on $(-1, 1) \times (0, \infty)$ which has been obtained by Eaton and Kariya (1977) and it is given by

$$\varphi_{\lambda,p}(t, r^2) = 2\frac{(\sqrt{\pi})^{p-1}}{\Gamma((p-1)/2)} r^{p-2}(1 - t^2)^{(p-3)/2} h(\sqrt{r^2 + \lambda^2 - 2\lambda r t}).$$

In any case, the monotone lilelihood property will hold if we can show that the derivative, with respect to t, of the expression $\log(h(\sqrt{r^2 + \lambda^2 - 2\lambda r t}))$ is nondecreasing in λ for all $t \in [-1, 1]$, $\lambda > 0$. We now return to the multivariate student distribution set up, that is h is given by $h(z) = (2\pi)^{-p/2}\mathrm{E}[V^{p/2} \exp(-z^2 V/2)]$ with $\mathcal{L}(V)$=Gamma$(d/2, d/2)$. We obtain

$$\begin{aligned}
\frac{\partial}{\partial t} \log(h(\sqrt{r^2 + \lambda^2 - 2\lambda r t})) &= \frac{\frac{\partial}{\partial t}\mathrm{E}[V^{p/2} \exp(-\{r^2 + \lambda^2 - 2\lambda r t\}V/2)]}{\mathrm{E}[V^{p/2} \exp(-\{r^2 + \lambda^2 - 2\lambda r t\}V/2)]} \\
&= \frac{\frac{\partial}{\partial t}\{r^2 + \lambda^2 - 2\lambda r t + d\}^{-(d+p)/2}}{\{r^2 + \lambda^2 - 2\lambda r t + d\}^{-(d+p)/2}} \\
&= \frac{(p + d)\lambda r}{(r^2 + \lambda^2 - 2\lambda r t + d)}
\end{aligned}$$

and the last expression is increasing in λ on $[0, \sqrt{d}]$ for all $r > 0$, $t \in [-1, 1]$.

□

Proof: expression (4) in Lemma 2. ¿From Lemma 1 we have

$$g_{h,\lambda}(r) = \lambda \frac{\int_0^\infty I_{\frac{p}{2}}(tv) v^{\frac{d}{2}} e^{-(\frac{d}{2}+s)v} dv}{\int_0^\infty I_{\frac{p}{2}-1}(tv) v^{\frac{d}{2}} e^{-(\frac{d}{2}+s)v} dv}$$
$$= \lambda \frac{\int_0^\infty I_{\frac{p}{2}}(x) x^{\frac{d}{2}} e^{-\frac{x}{u}} dx}{\int_0^\infty I_{\frac{p}{2}-1}(x) x^{\frac{d}{2}} e^{-\frac{x}{u}} dx}; \quad (5)$$

with the change of variables $x = tv$ and $u = 2\lambda r/(\lambda^2 + r^2 + d)$. Now, by expanding $I_\nu(x)$ and interchanging sum and integral, we obtain

$$\int_0^\infty I_\nu(x) x^{\frac{d}{2}} e^{-\frac{x}{u}} dx = \sum_{i\geq 0} \frac{(\frac{1}{2})^{2i+\nu}}{i!\Gamma(i+\nu+1)} \int_0^\infty x^{\nu+\frac{d}{2}+2i} e^{-\frac{x}{u}} dx$$
$$= (\frac{1}{2})^\nu \sum_{i\geq 0} \frac{(\frac{1}{2})^{2i}}{i!\Gamma(\nu+1)} \frac{\Gamma(\nu+\frac{d}{2}+1+2i) u^{\nu+\frac{d}{2}+2i+1}}{(\nu+1)_i}$$
$$= \frac{(\frac{1}{2})^\nu \, u^{\nu+\frac{d}{2}+1}}{\Gamma(\nu+1)} \sum_{i\geq 0} \frac{u^{2i}}{i!} \frac{\Gamma(\nu+\frac{d}{2}+1) \, (\frac{\nu+\frac{d}{2}+1}{2})_i \, (\frac{\nu+\frac{d}{2}+2}{2})_i}{(\nu+1)_i}$$
$$= \frac{\Gamma(\nu+\frac{d}{2}+1)}{\Gamma(\nu+1)} (\frac{1}{2})^\nu u^{\nu+\frac{d}{2}+1} \, {}_2F_1(\frac{\nu+\frac{d}{2}+1}{2}, \frac{\nu+\frac{d}{2}+2}{2}; \nu+1; u^2) \quad (6)$$

with ${}_2F_1(a_1, a_2; a_3; z) = \sum_{i\geq 0} \frac{(a_1)_i \, (a_2)_i}{(a_3)_i} \frac{z^i}{i!}$; and $(d)_i = \frac{\Gamma(d+i)}{\Gamma(d)}$. By using standard operations on hypergeometric functions, for any a_1, a_2, a_3 with $a_3 > 0$, we obtain that:

$${}_2F_1(a_1, a_2+1; a_3+1; z) = \sum_{i\geq 0} \frac{(a_1)_i \, (a_2+1)_i}{(a_3+1)_i} \frac{z^i}{i!}$$
$$= \sum_{i\geq 0} \frac{(a_1)_i \, (a_2)_i \frac{(a_2+i)}{a_2}}{(a_3)_i \frac{(a_3+i)}{a_3}} \frac{z^i}{i!}$$
$$= \frac{a_3}{a_2} \sum_{i\geq 0} \frac{(a_1)_i \, (a_2)_i}{(a_3)_i} \frac{(a_3+i)+(a_2-a_3)}{a_3+i} \frac{z^i}{i!}$$
$$= \frac{a_3}{a_2} [{}_2F_1(a_1, a_2; a_3; z) + \frac{a_2-a_3}{a_3} {}_2F_1(a_1, a_2; a_3+1; z)] \quad (7)$$

Combining the results (5), (6) and (7) with $u = 2\lambda r/(\lambda^2 + r^2 + d)$, $\nu = p/2$, $a_1 = (p+d+2)/4$, $a_2 = (p+d)/4$ and $a_3 = p/2$ we obtain:

$$g_{h,\lambda}(r) = \frac{\lambda^2 r}{\lambda^2 + r^2 + d} \{2 + \frac{(d-p)}{p} \mathrm{E}_u[\frac{p/2}{p/2+Y}]\}$$

where Y is a discrete random variable having probability mass function p_u with

$$p_z(y) \propto \frac{(a_1)_y (a_2)_y}{(a_3)_y} \frac{z^{2y}}{y!}; \; y = 0, 1, ... \quad (8)$$

for $0 < z < 1$. Since the family of distributions of Y has an increasing monotone likelihood ratio in Y with u viewed as the parameter, and the expression $(p/2+y)^{-1}$ is decreasing in y, it follows that $\mathrm{E}_u[\frac{p/2}{p/2+Y}]\}$ is decreasing in u but u is increasing in λ whenever $\lambda \leq \sqrt{d}$ so $\mathrm{E}_u[\frac{p/2}{p/2+Y}]\}$ is decreasing in λ. Since $0 \leq \mathrm{E}_u[\frac{p/2}{p/2+Y}]\} \leq 1$ we obtain that

$$g_{h,\lambda}(r) \leq \frac{\lambda^2 r}{\lambda^2 + r^2 + d}\{1 + (1 \vee \frac{d}{p})\}$$

for all λ, $0 \leq \lambda \leq m$. Setting $\lambda = m$ leads to the conclusion.

□

References

Cellier, D. & Fourdrinier, D. (1995). Shrinkage estimators under spherical symmetry for the general linear model. *Journal of Multivariate Analysis*, **52**, 338-351.

Kariya, T. & Eaton, M. (1977). Robust tests for spherical symmetry. *Annals of Statistics*, **5**, 206-215.

Marchand, É. & Perron, F. (2001). Improving on the MLE of a bounded normal mean. *Annals of Statistics*, **29**, 1066-1081.

Marchand, É. (1993). Estimation of a multivariate mean with constraints on the norm. *Canadian Journal of Statistics*, **21**, 359-366.

An Empirical Bayes Estimator for Weibull Distribution

Mohsen Mohammadzadeh

A11013

Department of Statistics, Tarbiat Modarres University, Iran.

Abstract. In empirical Bayes estimation of the parameter of continuous exponential family one usually uses estimators of marginal density of observations and its first derivative to approximate the Bayes estimator. In this paper the spline density estimation technique is used to estimate the marginal density and its derivative. Then an empirical Bayes estimator for the scale parameter of Weibull distribution is derived. Next the accuracy of this estimator is compared in a simulation study with two other estimators, a Bayes estimator with a Gamma prior distribution and an approximate Bayes estimator when the prior is a Gamma with unknown parameters.

Keywords. Empirical Bayes, Splines, Exponential Family

1 Introduction

Let X be a random variable of the one parameter natural exponential family with conditional density

$$f(x|\theta) = m(x)\exp\{T(x)\theta - A(\theta)\}, x \in R \tag{1}$$

where $\theta \in \Omega = \{\theta : \int \exp\{T(x)\theta - A(\theta)\}dx < \infty\}$, A is a real valued function of the parameter θ and T is a real valued statistic, so that its derivative with respect to x exists and is not zero. If θ is a realization of a variable Θ with a prior distribution G, then the marginal density of x is given by $f_G(x) = \int f(x|\theta)dG(\theta)$, and the Bayes estimator of θ, under the squared error loss, is the posterior mean of Θ, given by

$$\delta_G(x) = \frac{1}{T'(x)}\left[\frac{f'_G(x)}{f_G(x)} - \frac{m'(x)}{m(x)}\right] \tag{2}$$

Since in empirical Bayes methods, G is assumed to be unknown, $\delta_G(x)$ cannot be obtained. Suppose however, that we have n previous independent observations $x_1, \ldots, x_n$ from distribution with densities $f(x_1|\theta_1), \ldots, f(x_n|\theta_n)$ where $\theta_1, \ldots, \theta_n$ are independent realizations of the random variable Θ with distribution function G. These previous observations can be used to estimate $f_G(x)$ which in principle can be used to estimate G and hence to obtain an estimate of $\delta_G(x)$. This procedure can be complicated in general, but in the special case under consideration, i.e. in the case $f(x|\theta)$ is of the exponential family, $\delta_G(x)$ is of the form (2), it can be directly estimated from estimates of $f_G(x)$ and $f'_G(x)$.

A procedure that attempts to approximate the Bayes estimator (2) based on the kernel density estimate of the marginal density $f_G(x)$ and its derivative $f'_G(x)$ is used by Mohammadzadeh (2000) to find an empirical Bayes estimator for the unknown parameter of one parameter exponential family. In this paper, spline function

will be used to estimate f_G and f'_G. Then, using the current observation x together with these estimators, an empirical Bayes estimator for θ is derived. Next an approximate Bayes estimator, when the prior is a Gamma with unknown parameters, is given as another empirical Bayes estimator. Then two empirical Bayes estimators for the scale parameter of Weibull distribution are derived, and their accuracy is compared with a Bayes estimator in a simulation study.

The spline density estimators of f_G and f'_G are given in section 2. The Bayes and two empirical Bayes estimators for the scale parameter of Weibull distribution are derived in section 3. The simulation and results are described in section 4.

2 Spline Density Estimation

Spline density estimators with order $r \geq 2$ are introduced by Ciesielski (1991) and Krzykowski (1992). The spline density estimator for the unknown density function $f_G(x)$ and its derivative can be constructed by using the sample $X_1, \ldots, X_n$ and the current observation $X = x$. Namely, the r-spline density estimate of $f_G(x)$ is given by

$$f_n(x) = \sum_{s=s_0}^{s_m} a_s F_s(x), \tag{3}$$

and its derivative with respect to x is

$$f'_n(x) = \sum_{s=s_0}^{s_m} a_s F'_s(x), \tag{4}$$

where

$$s_0 = \left[\frac{X_{min}}{h} - \nu\right] - r, \qquad s_m = \left[\frac{X_{max}}{h} - \nu\right] + 1, \tag{5}$$

the constant ν is zero or $\frac{1}{2}$depending on r is even or odd, and $[y]$ denotes the integer part of y,

$$a_s = \frac{1}{nh} \sum_{j=1}^{n} F_s(X_j), \qquad s = s_0, s_0 + 1, \ldots, s_m,$$

and

$$F_s(x) = \sum_{k=1}^{r} N_{k,s}(x) I_{A_{k,s}}(x), \qquad F'_s(x) = \sum_{k=1}^{r} N'_{k,s}(x) I_{A_{k,s}}(x)$$

where, for $k = 1, \ldots, r$

$$A_{k,s} = [(s + \nu + k - 1)h, (s + \nu + k)h]$$

and

$$N_{k,s}(x) = \sum_{i=k}^{r} r \frac{(-1)^{r-i}}{i!(r-i)!} (s + \nu + i - \frac{x}{n})^{r-1},$$

$$N'_{k,s}(x) = \sum_{i=k}^{r} \frac{r(r-1)}{n} \frac{(-1)^{r-i+1}}{i!(r-i)!} (s + \nu + i - \frac{x}{n})^{r-2}.$$

An optimal value for h in (5) introduced by Krzykowski (1992, 1994) is $h = S\sqrt{\frac{6}{rn}}$, where S is the standard deviation of $X_1, \cdots, X_n$.

3 Estimation of Weibull Distribution Parameter

The Weibull distribution is widely used in reliability and life data analysis due to its versatility. Depending on the values of the parameters, this distribution can be used to model a variety of life behaviors. In this section we give three estimators for its scale parameter.

Suppose X has Weibull distribution with density

$$f(x;\theta) = 2\theta x \exp\{-\theta x^2\}, \qquad x > 0, \quad \theta > 0,$$

where θ is the scale parameter of this distribution. This density function can be written as an exponential family with $m(x) = 2x$, $T(x) = -x^2$ and $A(\theta) = -\log(\theta)$. According to (2), under the squared error loss, the Bayes estimator of θ becomes

$$\delta_G(x) = \frac{1}{2x}\left[\frac{1}{x} - \frac{f'_G(x)}{f_G(x)}\right]. \tag{6}$$

Based on the past observations of a given sample $X_1, \ldots, X_n$ and a current observation $X = x$ obtained from the distribution $f(x;\theta)$, the functions $f_G(x)$ and $f'_G(x)$ can be estimated using the spline density estimators (3) and (4). Then an empirical Bayes estimator of θ is given by

$$\delta_n(x) = \frac{1}{2x}\left[\frac{1}{x} - \frac{f'_n(x)}{f_n(x)}\right]. \tag{7}$$

Now suppose the prior distribution of Θ is Gamma(a, b). If a and b are known, the Bayes estimator of θ from (6) is given by

$$\delta_G(x) = \frac{a+n}{b + \sum_{i=1}^n x_i^2} \tag{8}$$

If the prior is a Gamma distribution with unknown parameters a and b, we can use the moment estimators

$$\hat{a} = \frac{(n-1)(\sum_{i=1}^n \frac{1}{x_i^2})^2}{n^2 \sum_{i=1}^n (\frac{1}{x_i^2} - \frac{1}{n}\sum_{i=1}^n \frac{1}{x_i^2})^2}$$

$$\hat{b} = \frac{(n-1)\sum_{i=1}^n \frac{1}{x_i^2}}{n \sum_{i=1}^n (\frac{1}{x_i^2} - \frac{1}{n}\sum_{i=1}^n \frac{1}{x_i^2})^2}$$

to define another empirical Bayes estimator for θ as

$$\delta'_n(x) = \frac{\hat{a}+n}{\hat{b} + \sum_{i=1}^n x_i^2} \tag{9}$$

n	δ_G	δ'_n	δ_n			
			r=2	r=3	r=4	r=5
10	.00402	.00968	.04778	.04583	.03816	.03265
15	.00439	.00464	.04792	.04405	.03727	.03009
30	.00334	.00359	.04779	.04289	.03665	.02704
50	.00254	.00341	.04034	.03908	.02681	.02634
80	.00234	.00239	.03847	.02813	.02541	.02526
100	.00229	.00233	.03601	.02798	.02317	.02241

Table 1. Table of Mean Square Errors .

4 Simulation and Results

In this section a simulation study has been done to compare the accuracies of the empirical Bayes estimators (7) and (9) and the Bayes estimator (8) in terms of their mean square errors (MSE). The simulation has been carried out according to the following scheme:

i) A value for θ is generated from distribution $Gamma(a, b)$ with $a = 2$ and $b = 3$.

ii) A random sample $x_1, \ldots, x_n, x$ with size $n + 1$ has been generated from distribution $Weibull(\theta)$. The Bayes estimator is computed from relation (6). The empirical Bayes estimator based on splines is computed from relation (7). Another empirical Bayes estimator as an approximated Bayes estimator with a Gamma prior and unknown parameters, using the moment estimators of a and b is computed from (9).

iii) For each value of $n = 10, 15, 30, 50, 80, 100$ and $r = 2, 3, 4, 5$ the above steps has been repeated 1000 times.

iv) The MSE associated with each estimator has been computed.

The MSE of 3 estimators for different values of n and r are summarized in table (1). The obtained values of MSE show that: the larger sample sizes give more accurate estimators and the Bayes estimator has always the smallest MSE, which are two obvious expected results. The empirical Bayes estimator $\delta'_n(x)$ has smaller MSE than $\delta_n(x)$. This is also an expected result, because it imparts from known form assumption of prior distribution. Increasing r, the degree of spline density estimator, gives interesting smaller MSE for $\delta'_n(x)$. But we should note that large values of r causes long computational times for empirical Bayes estimation. So a further study is needed for an optimal value of r. Therefore, the empirical Bayes estimator based on spline density estimation is well accurate and it can be used when the prior distribution is completely unknown. Although, for a more detailed study it is essential to compare this estimator with empirical Bayes estimator derived in terms of the kernel density estimators.

References

Ciesielski, Z. (1991), Asymptotic Nonparametric Spline Density Estimation, *Probability and Mathematical Statistics*, **12**, 1-24.

Krzykowski, G. (1992), Equivalent Conditions for the Nonparametric Spline Density Estimators, *Probability and Mathematical Statistics*, **13**, 269-276.

Krzykowski, G. (1994), The Choosing of the Window Parameter in Spline Density Estimation, *Preprint in the University of Gdansk*, Nr **98**.

Mohammadzadeh, M. (2000), Empirical Bayes Estimation for Contaminated Data, *Proceedings in Computational Statistics, COMPSTAT 2000*, Netherland.

Dependent Data, Moderate Deviations, and Density Estimation

Mojirsheibani, M.

P17006

Carlton University, Canada.

Abstract. In this presentation we consider some new bounds on the moderate deviations of strongly mixing sequences with applications to density estimation.
Keywords. Alpha-Mixing, Coupling, Convergence.

1 Introduction

Let $\mathbf{X}_1, \cdots, X_n$ be a strongly mixing sequence of zero-mean random variables. When X_i's are bounded, Hoeffding-type exponential inequalities are available for the large deviations of the partial sums $S_n = \sum_{i=1}^n$. In the unbounded case, under the so-called Cramer's condition ($E|X_i|^k \leq C_k E X_i^2 < \infty$, for positive constants C_k, and $k \geq$), one can establish Bernstein-type inequalities for the large deviations of S_n. More specifically, the following results are well-known (see, for example, Bosq (1998)):

Hoeffding-type exponential inequalities: *Let X_t be a zero-mean process. If* $\sup_t |X_t| \leq b$, *then for each* $q \in [1,\ n/2]$ *and each* $\epsilon > 0$,

$$P\{|S_n| > n\epsilon\} \leq 4\exp(\frac{-q\epsilon}{8b^2}) + 22(1 + 4b\epsilon^{-1})^{1/2} q\alpha([n/2q]). \qquad (*)$$

Here $\alpha(\cdot)$ is the mixing coefficient of the sequence (to be defined later at the beginning of section 2).

Bernstein-type inequalities: *Let X_t be a zero-mean process. Under the Cramer's condition (see above) one has for each* $q \in [1,\ n/2]$, $\epsilon > 0$, *and* $k \geq 3$

$$P\{|S_n| > n\epsilon\} \leq ((2n/q)+a)\exp(\frac{-q\epsilon^2}{25m_2^2 + 5c\epsilon}) + nb(k)\alpha([n/(q+1)])^{2k/(2k+1)}, \qquad (**)$$

where a, c, m_2^2, and $b(k)$ are positive constant not depending on n.

More recently, Liebscher (1996) used a result of Rio (1995) to establish the following bound:

Liebscher (1996): *Suppose that* $|X_i| \leq T(n) < \infty$. *Then for all* $n \geq 1$, *all integers* N *satisfying* $1 \leq N \leq n$, *and all* $\epsilon > 4NT(n)$,

$$P\{|S_n| > \epsilon\} \leq 4\exp\{-\epsilon^2(64nN^{-1}D(n,N) + 8\epsilon NT(n)/3)^{-1}\} + 4nN^{-1}\alpha(N), \qquad (***)$$

where $D(n,N) = \sup_{0\leq j\leq n-1} E(\sum_{i=j+1}^{(j+N)\wedge n})^2$.

The above bounds are used to study the almost sure behavior of S_n (LIL-type results). They are also used to study the rates of convergence in density estimation.

We will briefly give some results for density estimates. Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be a strongly mixing and strictly stationary sequence of $\Re^d$-valued random vectors. Let $f = F'$ be the density of $\mathbf{X}_i$. Let

$$f_n(\mathbf{x}) = (nh_n^d)^{-1} \sum_{i=1}^{n} K((\mathbf{x} - \mathbf{X}_i)/h_n)$$

be the usual kernel density estimate of f, where $K : \Re^d \to \Re$, called the kernel function, is typically required to satisfy certain regularity conditions. Here, the sequence h_n satisfies $h_n \to 0$, with $nh_n^d \to \infty$, as $n \to \infty$. When the kernel function K is uniformly bounded, then the Borel-Cantelli lemma in conjunction with (*) or (**) or (***) can provide rates of almost convergence of f_n to f, under various regularity conditions on K. Here we give one such result:

[Bosq (1998), Lemma 2.1]. *Let $\{X_t\}$ be strictly stationary and geometrically strongly mixing (i.e., $\alpha(k) \leq |const|\rho^k$, wher $k \geq 1$ and $0 \leq \rho < 1$) sequence of $\Re^d$-valued random vectors with density $f = F'$. If $h_n = c_n(\log n/n)^{1/(d+4)}$, where $c_n \to c > 0$, then under regularity conditions on f and the kernel, one has for all $\mathbf{x} \in \Re^d$ and all integers k*

$$(\mathrm{Log}_k n)^{-1}(n/\log n)^{2/(d+4)}\{f_n(\mathbf{x}) - f(\mathbf{x})\} \longrightarrow_{a.s.} 0.$$

Here $\mathrm{Log}_k n$ is the k-th iterated logarithm of n

In the rest of this article we present nonasymptotic bounds on the moderate deviations of S_n under the minimal assumption of the existence of a $2+c$ moments, for some $c > 0$. The resulting bounds can of course be used to establish rates of almost sure convergence for kernel density estimates under more relaxed conditions on the kernel function.

2. Main results. Let $\{X_t, t \in \mathcal{Z}\}$ be a strongly mixing process. That is,

$$\alpha(m) := \sup_{p \geq 1} \sup_{A \in \mathcal{F}_{-\infty}^{p},\ B \in \mathcal{F}_{p+m}^{\infty}} \left|P(A \cap B) - P(A)P(B)\right| \longrightarrow 0, \quad \text{as } m \to \infty.$$

Here, $\mathcal{F}_j^k = \sigma(X_i,\ j \leq i \leq k)$. Suppose that X_i and the mixing coefficient $\alpha(\cdot)$ satisfy the conditions

$$E|X_i|^{2+c} < \infty, \quad \text{for some } c > 0, \tag{1}$$

and

$$\alpha(N) \leq |const|N^{-m}, \quad \text{where } m > \frac{2+c}{c}. \tag{2}$$

Under (2), $\sum_{i \geq 1} \alpha(i)^{1-\frac{2}{2+c}} < \infty$, and hence $\sigma^2 := \sum_{i \in \mathcal{Z}} \mathrm{cov}(X_0, X_i) \geq 0$. We will also assume that $\sigma^2 > 0$. For any $\alpha \in (0,\ 1)$, put

$$n_\alpha := E(X_1 + \cdots + X_{[n^\alpha/2]})^2, \tag{3}$$

where [ ] is the usual greatest integer function. In what follows, for the ease of notation, we will take $\sigma^2 = 1$. (In fact, any $\sigma^2 > 0$ would be admissible.)

Theorem 1. *Let $\{X_t, t \in \mathcal{Z}\}$ be a strongly mixing and strictly stationary sequence of real-valued zero-mean random variables with a symmetric distribution (at mean) satisfying conditions* (1) *and* (2). *Then for every $\epsilon \in (0,\ c/2)$ and every $\alpha \in (0,\ 2\epsilon c^{-1} \wedge 0.5)$, and every m satisfying (as appears in (2))*

$$m > \max\Big\{\frac{5c+12(1-\alpha)}{8\alpha},\ \frac{2+c}{c},\ \frac{1-\epsilon+c/2}{\min(\alpha,\ \epsilon(4(2+c))^{-1}}-1,\ \frac{(c+2\alpha)(2+c)}{\alpha}-1\Big\}, \tag{4}$$

there are constants $b \in (0, \epsilon/2)$ and $n_o = n(c,\epsilon)$ such that for all $n > n_o$ and all real t_n, with

$$t_n^2 \ \geq\ 25(c+1)n^{-1}[n^{1-\alpha}]n_\alpha \log[n^{1-\alpha}], \tag{5}$$

one has

$$P\{|S_n| \geq \sqrt{n}\ t_n\} \ \leq\ C_1 n^{-\frac{c}{2}+\epsilon} t_n^{-2(c+2)} + C_2 n^{-\frac{c}{2}-\nu+\epsilon},\ \ \forall \nu < b.$$

Here, C_1 and C_2 are positive constants.

Theorem 2. *Let ϵ and α be as in Theorem 1. Under the conditions of Theorem 1, there is a $n_o = n(c,\epsilon)$ such that for all $n > n_o$ and all real t_n, with*

$$t_n^2 \ <\ 25(c+1)n^{-1}[n^{1-\alpha}]n_\alpha \log[n^{1-\alpha}], \tag{6}$$

one has

$$P\{|S_n| \geq \sqrt{n}\ t_n\} \ \leq\ C_3\ e^{-C(c,\alpha,\epsilon)(t_n^2 \wedge \log n)},$$

where the constant $C(c,\alpha,\epsilon)$ can be taken to be

$$0 < C(c,\alpha,\epsilon) = \min\left\{\frac{(c\wedge 1)(1-\alpha)}{2} + \frac{2(c+1)-(c\wedge 1)}{100(c+1)},\ \frac{c}{2}+\nu-\epsilon,\ 26^{-1}\right\}.$$

To prove theorems 1 and 2 we first need to state the following result of Bradley (1983).

Theorem 3. *Let $(Y_1, \cdots, Y_d,\ Y)$ be a vector in $\Re^{d+1}$. Suppose that $E|Y|^\gamma < \infty$ for some $\gamma > 1$. Let $\delta \in (0,\ (E|Y|^\gamma)^{1/\gamma}]$.* Then there is a random variable W such that
(a) $W =^d Y$,
(b) W and $(Y_1, \cdots, Y_d)$ are independent, and
(c) $P\{|W-Y| > \delta\} \leq 18\big(\frac{E|Y|^\gamma}{\delta^\gamma}\big)^{\frac{1}{2\gamma+1}} \Big(\sup_{A\in\sigma(Y_1,\cdots,Y_d),\ B\in\sigma(Y)} \big|P(A\cap B)-P(A)P(B)\big|\Big)^{\frac{2\gamma}{2\gamma+1}}$
The above theorem is often used as a coupling device to replace weakly dependent random variables with independent ones that have the same distribution.

PROOF OF THEOREM 1.
In what follows, C, $|const|$, C_o, $C_1, \cdots$ denote positive constants. Put $p = q = [n^\alpha/2]$ and $k = [n^{1-\alpha}]$ and define

$$W_i = X_{(i-1)(p+q)+1} + \cdots + X_{ip+(i-1)q},\ \ \text{for } 1 \leq i \leq k,$$

and

$$U_i = X_{ip+(i-1)q+1} + \cdots + X_{i(p+q)},\ \ \text{for } 1 \leq i \leq k.$$

Also, put

$$R(n) = X_{k(p+q)+1} + \cdots + X_n, \text{if } k(p+q) < n; \text{otherwise } R(n) = 0. \tag{7}$$

Now, repeated applications of Theorem 3 produces independent random variables $W_1^*, \cdots, W_k^*$, where $W_i^* =^d W_i$, and for any $\delta \in (0,\ (E|W_i|^2)^{1/2}]$,

$$P\{|W_i - W_i^*| > \delta\} \le 18(\delta^{-2}E|W_i|^2)^{1/5}(\alpha(q))^{4/5}. \tag{8}$$

Similarly, one can construct k random variables $U_1^*, \cdots, U_k^*$, where $U_i^* =^d U_i$, and for any $\delta \in (0,\ (E|U_i|^2)^{1/2}]$,

$$P\{|U_i - U_i^*| > \delta\} \le 18(\delta^{-2}E|U_i|^2)^{1/5}(\alpha(p))^{4/5}.$$

Now observe that

$$S_n \ = \ \sum_{i=1}^k W_i^* + \sum_{i=1}^k (W_i - W_i^*) + \sum_{i=1}^k U_i^* + \sum_{i=1}^k (U_i - U_i^*) + R(n). \tag{9}$$

To deal with the first term on the right side of (9), we need to state the following lemma:

Lemma 1. *Let $Y_1, \cdots, Y_n$ be iid random variables with $E(Y_1) = 0$, $E(Y_1^2) = 1$, and $E|Y_1|^{2+c} < \infty$, for some $c > 0$. Then there exist positive constants b and r, not depending on n, such that for all $n \ge 1$ and $t_n^2 \ge (c+1)\log n$,*

$$P\{\sum_{i=1}^n Y_i > \sqrt{n}\ t_n\} \le bn^{-c/2}t_n^{-2(c+2)} + nP\{|Y_1| > r\sqrt{n}\ t_n\}.$$

In fact, one may take $r = (2(c+1)(c+2))^{-1}$.

PROOF OF LEMMA 1.
The proof of this lemma is precisely that of the proof of Theorem 2 of Michel (1976) and will not be repeated here.

Since $E(W_i^*) = E(W_i) = 0$, and $E(W_i^*)^2 = E(W_i^2) = E(X_1 + \cdots + X_p)^2 =: n_\alpha$, where $p = [n^\alpha/2]$ and n_α is as in (3), one finds

$$E(W_i^*/\sqrt{n_\alpha}) = 0, \quad \text{and} \quad \mathrm{Var}(W_i^*/\sqrt{n_\alpha}) = 1.$$

Furthermore, for each p,

$$E\big|W_i^*/\sqrt{n_\alpha}\big|^{2+c} = (n_\alpha)^{-1-c/2}E|X_1+\cdots+X_p|^{2+c} \le p^{2+c}(n_\alpha)^{-1-c/2}E|X_1|^{2+c} < \infty.$$

Therefore for every $n \ge 1$ and t_n satisfying (22),

$$\begin{aligned} P\{\sum_{i=1}^k W_i^* > 5^{-1}\sqrt{n}\ t_n\} \ &= \ P\Big\{\sum_{i=1}^k \frac{W_i^*}{\sqrt{n_\alpha}} > \sqrt{k}\ t_n\big(5^{-1}\sqrt{n/(kn_\alpha)}\big)\Big\} \\ &\le \ C_4\ k^{-\frac{c}{2}}\big(5^{-1}t_n\sqrt{n/(kn_\alpha)}\ \big)^{-2(c+2)} \\ &\qquad +kP\Big\{\Big|\frac{W_1^*}{\sqrt{n_\alpha}}\Big| > C_5\sqrt{k}\ t_n\sqrt{n/(kn_\alpha)}\ \Big\} \\ &\qquad\qquad \text{(by Lemma 1 in conjunction with (22))} \\ &:= B_1(n) + B_2(n). \end{aligned} \tag{10}$$

Now with $k = [n^{1-\alpha}]$, $p = [n^{\alpha}/2]$, and the fact that $p/n_\alpha = 1 + o(1)$ one finds, for large n,

$$\begin{aligned} B_1(n) &\le C_6\ n^{-c(1-\alpha)/2} t_n^{-2(c+2)} \big(\frac{n}{kp}\big)^{-c-2} \big(\frac{p}{n_\alpha}\big)^{-c-2} \\ &\le C_7\ n^{\epsilon - c/2} t_n^{-2(c+2)} \big(\frac{p}{n_\alpha}\big)^{-c-2} n^{-\epsilon + \alpha c/2} \\ &\le C_8\ n^{\epsilon - c/2} t_n^{-2(c+2)}, \quad (\text{since } \alpha \le 2c^{-1}\epsilon). \end{aligned} \tag{11}$$

As for the term $B_2(n)$, since $W_1^* \ =^d W_1 \ = \ X_1 + \cdots + X_p \ := S_p$,

$$B_2(n) = kP\{|S_p| > C_5\ t_n\sqrt{n}\}. \tag{12}$$

Truncate X_i according to,

$$X_i' = X_i \mathbf{I}\big\{|X_i| \le n^{\frac{1}{2} - \frac{\epsilon}{2(2+c)}}\big\},$$

where $\mathbf{I}\{A\}$ denotes the indicator of the set A, and define $S_p' \ := X_1' + \cdots + X_p'$. Then, one obtains

$$\begin{aligned} P\{|S_p| > C_5\ t_n\sqrt{n}\} &\le P\{|S_p'| > C_5\ t_n\sqrt{n}\} + pP\{|X_1| > n^{\frac{1}{2} - \frac{\epsilon}{2(2+c)}}\} \\ &\le P\big\{|S_p'| > C_5\ t_n\sqrt{n}\big\} + C_9\ n^{\alpha - 1 - \frac{c}{2} + \frac{\epsilon}{2}} E|X_1|^{2+c} \qquad (13) \\ &\quad \text{(Markov's inequality and the fact that } p = [n^\alpha/2].) \qquad (14) \end{aligned}$$

To deal with the first term on the right side of (13), we need the following lemma whose proof will appear at the end of the section.

Lemma 2. *Let $\beta = \epsilon(2(2+c))^{-1}$. Also, let S_p' be as above. Then for every constant $d > 0$, and every $n > (2/d)^{1/(\beta - \min(\alpha,\ \beta/2))}$,*

$$P\{|S_p'| > d\ t_n\sqrt{n}\} \le 4\ e^{-C_9 n^{q(\alpha,\beta)}} + C_{10}\ n^{\alpha - (1+m)\min(\alpha,\ \beta/2)}.$$

Here, one may take $q(\alpha,\beta)$ to be

$$q(\alpha,\beta) = \frac{1}{2} - \frac{\max\big(\alpha,\ 1 - \min(\alpha,\ \beta/2)\big)}{2}\ (> 0).$$

Now observe tha Lemma 2 in conjunction with (12) and (13) imply that for large n,

$$\begin{aligned} n^{\frac{c}{2} - \epsilon} B_2(n) &= n^{\frac{c}{2} - \epsilon}[n^{1-\alpha}] P\{|S_p| > C_5\ t_n\sqrt{n}\} \\ &\le 4\ n^{\frac{c}{2} - \epsilon + 1 - \alpha} e^{-C_9 n^{q(\alpha,\beta)}} + C_{10}\ n^{\frac{c}{2} - \epsilon + 1 - (1+m)\min(\alpha,\ \beta/2)} + C_{11}\ n^{-\frac{\epsilon}{2}}, \end{aligned}$$

where $q(\alpha,\beta)$ is as in Lemma 2. Since by (23) $m > (1 + c/2 - \epsilon)\min^{-1}(\alpha,\ \beta/2) - 1$, one finds that $c/2 - \epsilon + 1 - (1+m)\min(\alpha,\ \beta/2) < 0$. Therefore, for every $\nu < (\epsilon/2) \wedge (-c/2 + \epsilon - 1 + (1+m)\min(\alpha,\ \beta/2))$,

$$B_2(n) \le C_{12}\ n^{-\frac{c}{2} - \nu + \epsilon}, \tag{15}$$

for n large enough. Putting together (10), (11), and (15), one concludes that for large n,

$$P\Big\{\sum_{i=1}^{k} W_i^* > 5^{-1}\sqrt{n}\ t_n\Big\} \le C_8\ n^{-\frac{c}{2}+\epsilon} t_n^{-2(c+2)} + C_{12}\ n^{-\frac{c}{2}-\nu+\epsilon}. \tag{16}$$

Similarly, since $p = q = [n^\alpha/2]$, one also finds for large n,

$$P\Big\{\sum_{i=1}^{k} U_i^* > 5^{-1}\sqrt{n}\ t_n\Big\} \le C_8\ n^{-\frac{c}{2}+\epsilon} t_n^{-2(c+2)} + C_{12}\ n^{-\frac{c}{2}-\nu+\epsilon}. \tag{17}$$

To deal with the term $\sum_{i=1}^{k}(W_i - W_i^*)$ in (9), first note that $E|W_i|^2 = p + o(p)$. Now, put

$$b^2(n) = 25(c+1)n^{-1}[n^{1-\alpha}]n_\alpha \log[n^{1-\alpha}]$$

and observe that that $b^2(n) = (25/2)(c+1)(1-\alpha)\log n + o(1)$. Combining this with the fact that $k = [n^{1-\alpha}] > n^{1-\alpha}/2$, one obtains, for large n (recall that $\alpha < 0.5$)

$$\frac{\sqrt{n}\ b(n)}{5k} \le |const|\ n^{\alpha-1/2}\sqrt{\log n + o(1)} \in \Big(0,\ (E|W_1|^2)^{1/2}\Big].$$

Therefore,

$$\begin{aligned}
P\big\{|W_i - W_i^*| > \sqrt{n}\ t_n/(5k)\big\} &\le P\big\{|W_i - W_i^*| > \sqrt{n}\ b(n)/(5k)\big\}, \quad \text{(by (22))} \\
&\le C_{12}\Big(\{\sqrt{n}\ b(n)/(5k)\}^{-2}\{C_{13}\ n^\alpha + o(n^\alpha)\}\Big)^{\frac{1}{5}} (\alpha(q))^{\frac{4}{5}} \\
&\quad \text{(by (8))} \\
&\le C_{14}\Big(n^{-1+2(1-\alpha)}\{C_{15}\ \log n + o(1)\}^{-1}\{C_{13}\ n^\alpha + o(n^\alpha)\}\Big)^{\frac{1}{5}} \\
&\quad \times\big(\alpha([n^\alpha/2])\big)^{\frac{4}{5}} \\
&\quad \text{(since } b^2(n) = (25/2)(c+1)(1-\alpha)\log n + o(1)) \\
&\le C_{16}\{n^{1-\alpha} + o(n^{1-\alpha})\}^{\frac{1}{5}}\{\log n + o(1)\}^{\frac{-1}{5}}\big([n^\alpha/2]\big)^{\frac{-4m}{5}} \\
&\quad \text{(by (2))} \\
&\le C_{17}\{n^{1-\alpha-4m\alpha} + o(n^{1-\alpha-4m\alpha})\}^{\frac{1}{5}}\{\log n + o(1)\}^{\frac{-1}{5}}.
\end{aligned}$$

Therefore

$$\begin{aligned}
P\Big\{\sum_{i=1}^{k}|W_i - W_i^*| > \sqrt{n}\ t_n/5\Big\} &\le C_{17}\ k\{n^{1-\alpha-4m\alpha} + o(n^{1-\alpha-4m\alpha})\}^{\frac{1}{5}}\{\log n + o(1)\}^{\frac{-1}{5}} \\
&\le C_{18}\{n^{6(1-\alpha)-4m\alpha} + o(6(n^{1-\alpha})-4m\alpha)\}^{\frac{1}{5}}\{\log n + o(1)\}^{\frac{-1}{5}} \\
&\le C_{18}\{n^{-5c/2} + o(n^{-5c/2})\}^{\frac{1}{5}}\{\log n + o(1)\}^{\frac{-1}{5}} \\
&\quad \text{(since } m > (5c + 12(1-\alpha))/(8\alpha), \text{ by (23))} \\
&< C_{18}\ n^{-\frac{c}{2}}.
\end{aligned} \tag{18}$$

Similarly, one can show that

$$P\Big\{\sum_{i=1}^{k}|U_i - U_i^*| > \sqrt{n}\ t_n/5\Big\} < C_{19}\ n^{-\frac{c}{2}}. \tag{19}$$

Finally, the term $R(n)$ that appears in (7) may be handled as follows. Define the truncated variables X_i'' by

$$X_i'' = X_i \mathbf{I}\big\{|X_i| \le n^{\frac{1}{2}-\frac{\min(\alpha,\ 1-\alpha)}{2+c}}\big\}$$

and put

$$R''(n) = \sum_{i=2[n^{1-\alpha}][n^{\alpha}/2]+1}^{n} X_i'' \ , \quad \text{if } \ 2[n^{1-\alpha}][n^{\alpha}/2] < n; \quad (\text{otherwise } \ R''(n) = 0).$$

Let $C_\alpha = n^\alpha/2 - [n^\alpha/2]$ and $C_{1-\alpha} = n^{1-\alpha} - [n^{1-\alpha}]$, and observe that the number of terms in the sum $R''(n)$ is at most

$$n - 2[n^{1-\alpha}][n^{\alpha}/2] \ = \ 2C_\alpha n^{1-\alpha} + C_{1-\alpha}n^{\alpha} - 2C_\alpha C_{1-\alpha} \ < \ 3n^{\max(\alpha,\ 1-\alpha)}.$$

Using Lemma 2, it is not difficult to show that

$$P\{|R''(n)| > \sqrt{n}\ t_n/5\} \ \le \ 4\ e^{-C_{20}\ n^{\phi(\alpha,c)}} + C_{21}\ n^{-\varphi(\alpha,m,c)}, \quad \forall\ n > 10^{\frac{2(2+c)}{\min(\alpha,\ 1-\alpha)}}, \tag{20}$$

with

$$2\phi(\alpha,\ c) \ = \ 1 - \max\left(\alpha,\ 1 - \min\left(\alpha,\ \frac{\min(\alpha,\ 1-\alpha)}{2(2+c)}\right)\right) \ = \ \frac{\alpha}{2(2+c)},$$

and

$$-\varphi(\alpha,\ m,\ c) \ = \ \alpha - (1+m)\min\left(\alpha, \frac{\min(\alpha,\ 1-\alpha)}{2(2+c)}\right) \ < \ -\frac{c}{2},$$

where the bound $-c/2$ in the above expression follows from the fact that $m > (c+2\alpha)(2+c)(\min(\alpha,\ 1-\alpha))^{-1} - 1$; see (23). Therefore, for $n > 10^{2(2+c)/\min(\alpha,\ 1-\alpha)}$,

$$\begin{aligned}
&P\{|R(n)| > \sqrt{n}\ t_n/5\} \\
&\le \ P\{|R''(n)| > \sqrt{n}\ t_n/5\} + \big(n - 2[n^{1-\alpha}][n^{\alpha}/2]\big)P\left\{|X_1''| > n^{\frac{1}{2}-\frac{\min(\alpha,\ 1-\alpha)}{2+c}}\right\} \\
&\le \ 4\ e^{-C_{20}\ n^{\phi(\alpha,c)}} + C_{21}\ n^{-\varphi(\alpha,m,c)} + 3\ n^{\max(\alpha,\ 1-\alpha)} \cdot n^{-1-\frac{c}{2}+\min(\alpha,\ 1-\alpha)}\ E|X_1''|^{2+c} \\
&\quad \text{(by (20) and Markov's inequality)} \\
&\le \ C_{22}\ n^{-\frac{c}{2}}.
\end{aligned} \tag{21}$$

Now, (16), (17), (18), (19), and (21) imply that for n large enough,

$$P\{S_n > \sqrt{n}\ t_n\} \ \le \ C_{23}\ n^{-\frac{c}{2}+\epsilon} t_n^{-2(c+2)} + C_{24}\ n^{-\frac{c}{2}-\nu+\epsilon}.$$

Similarly, (recall that $E(S_n) = 0$), the same bound holds for $P\{S_n < -\sqrt{n}\ t_n\}$. This completes the proof of Theorem 1. □

A more general version of Theorem 1 (without the assumption of a symmetric distribution) was proved by Mojirsheibani (2002). It may be stated as:

Theorem 4. *Let $\{X_t, t \in \mathcal{Z}\}$ be a strongly mixing and stationary sequence of real-valued zero-mean random variables satisfying conditions* (1) *and* (2). *Then for every $\epsilon \in (0,\ c/2]$ and every $\alpha \in (0,\ 2\epsilon c^{-1} \wedge 0.5)$ there are constants $m_o = m(c,\epsilon)$, $b \in (0, \epsilon/2)$, and $n_o = n(c,\epsilon)$ such that for all $n > n_o$, all $m > m_o$, and all real t_n, with*

$$t_n^2 \ \geq\ 8(c+1)\log[n^{1-\alpha}], \tag{22}$$

one has

$$P\{|S_n| \geq \sqrt{n}\ t_n\} \ \leq\ C_1 n^{-\frac{c}{2}+\epsilon} t_n^{-2(c+2)} + C_2 n^{-\frac{c}{2}-\nu+\epsilon},\ \ \forall \nu < b.$$

Here, C_1 and C_2 are positive constants, $S_n = X_1 + \cdots + X_n$, and the constant m_0 may be taken to be

$$m_o = \max\Big\{\frac{5c+12(1-\alpha)}{8\alpha},\ \frac{2+c}{c},\ \frac{1-\epsilon+c/2}{\min(\alpha,\ \epsilon(4(2+c))^{-1})} - 1\Big\}. \tag{23}$$

The proof of Theorem 4 is based a more careful application of Bradley's (1983) Theorem that leaves no remainder terms.

PROOF OF LEMMA 2.
Since $|X_i'|$ is bounded by $n^{\frac{1}{2}-\beta}$, where $\beta = \epsilon(2(2+c))^{-1}$ and has mean zero (the distribution of $X_i's$ is symmetric at the mean), Theorem 2.1 of Liebscher (1996) implies that for every integer N, with $1 \leq N \leq p\ := [n^\alpha/2]$,

$$P\{|S_p'| > d\ t_n\sqrt{n}\} \leq 4\ \exp\left\{\frac{-d^2 n t_n^2}{64N^{-1}pD(p,N) + 3^{-1}(8dt_n\sqrt{n})Nn^{0.5-\beta}}\right\} + \frac{4p}{N}\alpha(N), \tag{24}$$

provided that

$$d\ t_n\sqrt{n} > 4Nn^{0.5-\beta}. \tag{25}$$

Here, $D(p,N) = \sup_{0\leq j\leq p-1} E(\sum_{i=j+1}^{(j+N)\wedge p} X_i')^2$. Put $\ell = 2m/(m-1)$ and note that $\ell < 2+c$, (because $m > (2+c)/c$, by (23)). Now (2), Lemma 2.1 of Liebscher (1996), and the fact that $E|X_i'|^\ell \leq E|X_i|^\ell < (E|X_i|^{2+c})^{\ell/(2+c)} < \infty$, ($E|X_i'|$ does not depend on n), imply that

$$D(p,N) \leq |const|N\log p(E|X_i'|^\ell)^{2/\ell} = |const|N\log p \quad (E|X_i'| \text{ does not depend on } n)$$

Furthermore, choosing $N = [2^{-1}n^{\min(\alpha,\ \beta/2)}]$, it is not difficult to verify that (25) is satisfied for $n > (2/d)^{1/(\beta-\min(\alpha,\ \beta/2))}$. Thus, the above choice of N together with (24) imply that

$$\begin{aligned}
P\{|S_p'| > d\ t_n\sqrt{n}\} \leq 4\ \exp&\left\{\frac{-nt_n^2}{C_{25}\ p\log p + C_{26}\ t_n\sqrt{n}[2^{-1}n^{\min(\alpha,\ \beta/2)}]n^{0.5-\beta}}\right\} \\
&+\frac{4p}{[2^{-1}n^{\min(\alpha,\ \beta/2)}]}\alpha([2^{-1}n^{\min(\alpha,\ \beta/2)}]) \\
\leq 4\ \exp&\left\{\frac{-nt_n^2}{C_{27}\ n^\alpha\log n^\alpha + C_{28}\ n^{1+\min(\alpha,\ \beta/2)-\beta}t_n}\right\} \\
&+C_{29}\ n^{\alpha-\min(\alpha,\ \beta/2)}([2^{-1}n^{\min(\alpha,\ \beta/2)}])^{-m} \\
&\text{(by (2) and the fact that } [y] \geq y/2 \text{ for } y \geq 1)
\end{aligned}$$

$$= 4\ \exp\left\{ \frac{-n^{1-\psi(\alpha,\beta)}h(n)t_n^2}{n^{-\psi(\alpha,\beta)}h(n)\big(C_{27}\ n^{\alpha}\log n^{\alpha} + C_{28}\ n^{1+\min(\alpha,\ \beta/2)-\beta}t_n\big)} \right\}$$
$$+C_{30}\ n^{\alpha-(m+1)\min(\alpha,\ \beta/2)},$$

where

$$\psi(\alpha,\beta) = \max(\alpha,\ 1-\min(\alpha,\ \beta/2)),\ \text{ and }\ h(n) = \begin{cases} t_n^{-2} & \text{under (22)},\\ (\log n)^{-1} & \text{under (6)}. \end{cases}$$

Observe that $t_n^{-2}\log n \le C_{31}\ (1+o(1))$, under (22), where as $t_n^2(\log n)^{-1} < C_{32}\ (1+o(1))$ under (6). Since $\psi(\alpha,\beta) \ge 1-\beta+\min(\alpha,\ \beta/2) > 0$, it is straightforward to see that for large n, under (22),

$$P\{|S_p'| > d\ t_n\sqrt{n}\} \le 4\ e^{-C_{33}n^{1-\psi(\alpha,\beta)}} + C_{34}\ n^{\alpha-(1+m)\min(\alpha,\ \beta/2)}.$$

Similarly, under (6), one finds for large n,

$$P\{|S_p'| > d\ t_n\sqrt{n}\} \le 4\ e^{-C_{34}n^{(1-\psi(\alpha,\beta))/2}} + C_{35}\ n^{\alpha-(1+m)\min(\alpha,\ \beta/2)}.$$

This completes the proof of Lemma 2. □

PROOF OF THEOREM 2.
We first need the following lemma, (which is the counterpart of Lemma 1 under the condition that $t_n^2 \le (c+1)\log n$.

Lemma 3. *Let $Y_1,\cdots,Y_n$ be iid random variables with $E(Y_1)=0$, $E(Y_1^2)=1$, and $E|Y_1|^{2+c} < \infty$, for some $c>0$. Then there exist positive constants b and r, not depending on n, such that for all $n \ge 1$ and $t_n^2 \le (c+1)\log n$,*

$$\big|P\{\sum_{i=1}^{n} Y_i > \sqrt{n}\ t_n\} - \Phi(-t_n)\big| \le bn^{-\frac{c\wedge 1}{2}}e^{-(1-\sigma)t_n^2/2} + nP\{|Y_1| > r\sqrt{n}\ t_n\},$$

where $\sigma = 2^{-1}(c+1)^{-1}(c\wedge 1)$, and Φ is the standard normal distribution function.

PROOF OF LEMMA 3.
This is simply Theorem 1 of Michel (1976).

First observe that under (6), one has $t_n^2 25^{-1}(n/(kn_\alpha)) < (c+1)\log k$, where $k = [n^{1-\alpha}]$ as before. Combining this with Lemma 3, one can bound the term $\sum_{i=1}^{k} W_i^*$, that appears in the representation (9) of S_n, as follows:

$$P\{\sum_{i=1}^{k} W_i^* > 5^{-1}\sqrt{n}\ t_n\} = P\Big\{\sum_{i=1}^{k}\frac{W_i^*}{\sqrt{n_\alpha}} > \sqrt{k}\ t_n\big(5^{-1}\sqrt{n/(kn_\alpha)}\big)\Big\}$$
$$\le C_{36}\ k^{-\frac{c\wedge 1}{2}}\exp\left\{ -\frac{2(c+1)-(c\wedge 1)}{4(c+1)}\big(t_n^2(n/(25kn_\alpha))\big)\right\}$$
$$+kP\Big\{|W_i^*| > C_{37}\ t_n\sqrt{n}\ \Big\} + \Phi\Big\{ -t_n 5^{-1}\sqrt{n/(kn_\alpha)}\ \Big\}$$
$$:= B_3(n) + B_4(n) + B_5(n) \qquad (26)$$

The term $B_4(n)$ is the same as $B_2(n)$ in (10), except for possibly different constants. Therefore the argument leading to (15) gives the bound $B_4(n) \leq C'_{12}\ n^{-\frac{c}{2}-\nu+\epsilon}$, for some positive constant C'_{12} and all ν: $\nu < (\epsilon/2) \wedge (-c/2+\epsilon-1+(1+m)\min(\alpha,\ \beta/2))$. Next, since $[n^\alpha/2]/n_\alpha = 1 + o(1)$, where n_α is as in (3), we may write

$$\frac{n}{kn_\alpha} = \frac{n}{[n^{1-\alpha}][n^\alpha/2]} \times \frac{[n^\alpha/2]}{n_\alpha} = (2+o(1))(1+o(1)) = 2 + o(1) \tag{27}$$

Consequently, for large n we have

$$B_3(n) \leq C_{38}\ n^{-(c\wedge 1)(1-\alpha)/2} \exp\Big\{ -\frac{2(c+1)-(c\wedge 1)}{100(c+1)}\ t_n^2 \Big\}. \tag{28}$$

Finally, since $P\{N(0,1)\text{ r.v. } > x\} < x^{-1}(2\pi)^{-1/2}\exp(-x^2/2)$, for all $x > 0$, (Mill's ratio), one finds, in conjunction with (27), that for n large enough $B_5(n) < C_{39}\ t_n^{-1}e^{-t_n^2/26}$, where $C_{39} > 0$. Putting all the above together, for n large enough we have the bound:

$$P\{\sum_{i=1}^k W_i^* > 5^{-1}\sqrt{n}\ t_n\} \leq C_{40}\ e^{-C_{41}\ (t_n^2 \wedge \log n)}, \tag{29}$$

where

$$C_{41} = C(c,\alpha,\epsilon) = \min\Big\{ \frac{(c\wedge 1)(1-\alpha)}{2} + \frac{2(c+1)-(c\wedge 1)}{100(c+1)},\ \frac{c}{2}+\nu-\epsilon,\ 26^{-1} \Big\}.$$

Similarly (since $p = q = [n^\alpha/2]$), one can show that $P\{\sum_{i=1}^k U_i^* > 5^{-1}\sqrt{n}\ t_n\} \leq C_{40}\ \exp\{-C_{41}\ (t_n^2 \wedge \log n)\}$. Also, note that (6) in conjunction with (27) implies that for large n

$$\frac{\sqrt{n}\ t_n}{5k} < \frac{2n^{\alpha-1/2}\sqrt{|const|\log n + o(1)}}{5} \in \Big(0,\ (E|W_1|^2)^{1/2}\Big].$$

Now it is straightforward to show that for n large enough, (8) and the argument leading to (18) give the bound

$$P\Big\{\sum_{i=1}^k |W_i - W_i^*| > \sqrt{n}\ t_n/5\Big\} \leq C_{42}\ n^{-c/2}t_n^{-2/5}.$$

Similarly, by symmetry, the same bound holds for $P\Big\{\sum_{i=1}^k |U_i - U_i^*| > \sqrt{n}\ t_n/5\Big\}$. Combining these bounds with (18), which holds under both (22) and (6), we may conclude that under the conditions of Theorem 2,

$$P\{S_n > \sqrt{n}\ t_n\} \leq C_{43}\ n^{-c/2}t_n^{-2/5},$$

for a positive constant C_{102}. Also, the same bound holds for $P\{S_n < -\sqrt{n}\ t_n\}$, (recall that $E(S_n) = 0$). This completes the proof of Theorem 2. □

References

Bradley, R. 1983. Approximation theorems for strongly mixing random variables. *Michigan Math. J.* **30**, 69-81.

Bosq, D. 1998. Nonparametric Statistics for Stochastic Processes. Springer.

Liebscher, E. (1996). Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stoch. Process. Appl.* **65**, 69-80.

Michel, R. 1976. Nonuniform central limit bounds with applications to probabilities of deviation. *Ann. Probab.* **4**, 102-106.

Mojirsheibani, M. 2002. Some results on sums of unbounded mixing random variables. *Proceedings of the International Conference on Asymptotic Methods in Stochastics*. Fields Institute Publications. (To appear.)

Rio, E. (1995). The functional law of the iterated logarithm for stationary strongly mixing sequences. *Ann. Probab.* **23**, 1188-1203.

Asymptotic Equivalence of the Covariance Matrix of a Multivariate Stationary Time Series with the Related Circular Symmetric Matrix

A. R. Nematollahi and Z. Shishebor

P11004

Department of Statistics, Shiraz University, Iran

Spectral density function has fundamental role in the spectral domain, so its estimation is of interest. Subba Rao and Gabr have derived an estimation for the spectral density function of a stationary time series using the properties of the eigevalues of variance-covariance matrix. Nematollahi and Subba Rao extended this to multivariate case. They conjectured the asymptotic equivalence of the variance-covariance matrix with the related circular symmetric matrix.

In this paper we prove the asymptotic equivalence of the variance-covariance matrix of a multivariate stationary time series with the related circular symmetric matrix. The method is illustrated with simulated time series.

Keywords: spectral density matrix, eigenvalue decomposition, block-Toeplitz matrix.

1 Introduction

Much of the noticeable progress in applied sciences is under the effect of communication and electronic sciences. For example, in technology of satellite, radar, sonar, internet and etc. These progress can not be obtained without using time series analysis and signal processing. In time series analysis, there are two approaches for analysis, the first is time domain and the second is spectral (or frequency) domain. Although, there is a one to one corresponding between them, but in applied sciences, analysis in spectral domain are widely used, for more highlighting the information that is maybe hidden within the data.

Spectral density function has fundamental role in the spectral domain, so its estimation is of interest.

some important information about trend, seasonality, hidden periodicity and delaying time in linear systems can be derived by the spectral density function.

Subba Rao and Gabr (1988) derived an estimation for the spectral density function of a stationary time series using the properties of the eigenvalues of variance-covariance matrix. Nematollahi and Subba Rao (2002) extended this to multivariate case. They conjectured the asymptotic equivalence of the variance-covariance matrix with the related circular symmetric matrix. In this paper we prove this conjecture.

2 Preliminaries:

Let $\mathbf{Y} = \{\mathbf{Y}_t, t \in Z\}$ be a T-dimensional stationary real-valued time series with $E\mathbf{Y}_t = 0, E|Y_t(l)|^2 < \infty, l = 1, \ldots, T$ where $Y_t(l)$ is the l-th element of $\mathbf{Y}_t$ for all $t \in Z$, and Z stands for all integer numbers. Let $R(\tau) = E\mathbf{Y}_{t+\tau}\mathbf{Y}_t^{'}, t, \tau \in Z$ be $T \times T$ autocovariance matrix of $\mathbf{Y}_t$ (the symbol $'$ is denoted for transpose). We assume that $\mathbf{Y}_t$ has an absolutely continuous spectrum (spectral density)

$$\mathbf{h}(w) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \mathbf{R}(\tau) e^{-iw\tau}, \qquad 0 \leq w \leq 2\pi. \tag{2.1}$$

Let $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n\}$ be a random sample of size n from $\mathbf{Y}$, and consider a $nT \times 1$ vector $\mathcal{Y}_n = (\mathbf{Y}_n^{'}, \mathbf{Y}_{n-1}^{'}, \ldots, \mathbf{Y}_1^{'})^{'}$. The covariance matrix $\Gamma_n = E\mathcal{Y}_n\mathcal{Y}_n^{'}$ has the following form

$$\Gamma_n = \begin{pmatrix} \mathbf{R}(0) & \mathbf{R}(1) & \ldots & \mathbf{R}(n-1) \\ \mathbf{R}(-1) & \mathbf{R}(0) & \ldots & \mathbf{R}(n-2) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{R}(-(n-1)) & \mathbf{R}(-(n-2)) & \ldots & \mathbf{R}(0) \end{pmatrix}. \tag{2.2}$$

Note that Γ_n is a block-Toeplitz $nT \times nT$ matrix. Individual matrix elements are not ingeneral symmetric, $\mathbf{R}(\tau) \neq \mathbf{R}^{'}(\tau)$, although $\mathbf{R}(-\tau) = \mathbf{R}^{'}(\tau)$.

These type of matrices have too many applications in the time series analysis (see Nematollahi and Subba Rao, (2000), Hannan (1970), Hannan and Wahlberg (1989)).

3 Main Result

In this section we obtain a special matrix that will approximately diagonalize Γ_n. Hannan and Wahlberg, 1989, have obtained the following useful eigen-value decomposition of Γ_n,

$$\Gamma_n = \tilde{\mathbf{W}}_n^* \mathbf{D}_n \tilde{\mathbf{W}}_n, \tag{3.1}$$

where $\tilde{\mathbf{W}}_n$ is the $nT \times nT$ matrix with jth (block) row $\mathbf{W}_n(w_j)$ given by

$$\mathbf{W}_n(w_j) = n^{-1/2}(\mathbf{I}, e^{iw_j}\mathbf{I}, \ldots, e^{(n-1)iw_j}\mathbf{I}) \tag{3.2}$$

and $\mathbf{D}_n = diag\{\Lambda_n(w_0), \Lambda_n(w_1), \ldots, \Lambda_n(w_{n-1})\}$, with $\Lambda_n(w_j)$ is a $T \times T$ Hermitian matrix given by

$$\Lambda_n(w_j) = \mathbf{W}_n(w_j)\Gamma_n\mathbf{W}_n^*(w_j) \qquad j = 0, \ldots, n-1, \tag{3.3}$$

where $\mathbf{I}$ is an $T \times T$ identity matrix and $w_j = \frac{2\pi j}{n}, j = 0, \ldots, n-1$ be the frequencies. Note that $\tilde{\mathbf{W}}_n$ is an orthogonal matrix. The $\Lambda_n(w_j)$ is identified as an eigen-value matrix and$\mathbf{W}_n(w_j)$ is an eigen-vector matrix associated with $\Lambda_n(w_j)$ (Nematollahi and Subba Rao, 2000).

Now consider the following circular block matrix (circulant),

$$\Gamma_n^c = \begin{pmatrix} \mathbf{R}(0) & \mathbf{R}(1) & \dots \mathbf{R}(n-1) \\ \mathbf{R}(n-1) & \mathbf{R}(0) & \dots \mathbf{R}(n-2) \\ \mathbf{R}(n-1) & \mathbf{R}(n-1) & \cdots \mathbf{R}(n-3) \\ \vdots & \vdots & \vdots \\ \mathbf{R}(1) & \mathbf{R}(2) & \dots \mathbf{R}(0) \end{pmatrix}.$$

Setting $\mathbf{R}(1) = \mathbf{R}^{'}(n-1), \dots, \mathbf{R}(h) = \mathbf{R}^{'}(n-h)$ in Γ_n^c, we obtain the circular block Hermitian (indeed, symmetric) matrix

$$\Gamma_n^s = \begin{pmatrix} \mathbf{R}(0) & \mathbf{R}(1) & \mathbf{R}(2) \dots \mathbf{R}^{'}(2) & \mathbf{R}^{'}(1) \\ \mathbf{R}^{'}(1) & \mathbf{R}(0) & \mathbf{R}(1) \dots \mathbf{R}^{'}(3) & \mathbf{R}^{'}(2) \\ \mathbf{R}^{'}(2) & \mathbf{R}^{'}(1) & \mathbf{R}(0) \dots \mathbf{R}(4) & \mathbf{R}(3) \\ \vdots & \vdots & \vdots \quad \vdots \quad \vdots & \vdots \\ \mathbf{R}(2) & \mathbf{R}(3) & \mathbf{R}(4) \dots \mathbf{R}(0) & \mathbf{R}(1) \\ \mathbf{R}(1) & \mathbf{R}(2) & \mathbf{R}(3) \dots \mathbf{R}^{'}(1) & \mathbf{R}(0) \end{pmatrix}. \tag{3.4}$$

Each eigen-value matrix and eigen-vector matrix of Γ_n^s are denoted by $\Lambda_n^s(w_j)$ and $\mathbf{W}_n^s(w_j)$, respectively.

Before considering the main theorem, we need to express the following Lemmas:

Lemma 3.1. The eigen-value matrices and eigen-vector matrices of Γ_n^s in (3.4) satisfy

i) $\Lambda_n^s(w_j) = \Lambda_n^{'s}(w_{n-j})$.
ii) $\mathbf{W}_n^s(w_0) = n^{-1/2}(\mathbf{I}, \dots, \mathbf{I})$,
$\mathbf{W}_n^s(w_j) = n^{-1/2}2^{1/2}(\mathbf{0}, sinw_j\mathbf{I}, \dots, sin(n-1)w_j\mathbf{I})$,
$\mathbf{W}_n^s(w_{n-j}) = n^{-1/2}2^{1/2}(\mathbf{I}, cosw_j\mathbf{I}, \dots, cos(n-1)w_j\mathbf{I})$,

for $j = 0, 1, \dots, \frac{n-1}{2}$.

Proof. The proof of (i) follows from symmetry of Γ_n^s and relations (3.3) and (3.2).
We derive (ii) from (3.3) and part (i). Let $\mathbf{h}_n(w)$ be the "truncated spectral density matrix" ,i.e.,

$$\mathbf{h}_n(w) = \frac{1}{2\pi n} \sum_{t=1}^{n} \sum_{s=1}^{n} \mathbf{R}(t-s) e^{-i(t-s)w}, \tag{3.5}$$

(See Subba Rao and Gabr, 1989).

Lemma 3.2. The eigen-value matrices of Γ_n^s are approximately equal to the spectral density matrix of $\mathbf{Y}$ at n frequency points $w_j, j = 0, \dots, n-1$. More precisely, $\Lambda_n^s(w_j)$ converges to $2\pi\mathbf{h}(w_j)$ element by element.

Proof. The truncated spectral density matrix $\mathbf{h}_n(w)$ in (3.5) is asymptotically equivalent to $\mathbf{h}(w)$, under some regularity condition, (Priestley, 1981, P. 418). On the other hand, using (3.2) and (3.3) one can easily show that $\Lambda_n^s(w_j) = 2\pi\mathbf{h}_n(w_j)$.

Now define the orthogonal matrix $\mathbf{Q} = n^{-1/2}2^{1/2}\mathbf{B}^{'}$ where

$$\mathbf{B} = \begin{pmatrix} 2^{-1/2}\mathbf{I} & 2^{-1/2}\mathbf{I} & 2^{-1/2}\mathbf{I} & \cdots & 2^{-1/2}\mathbf{I} \\ \mathbf{I} & \cos 2\pi\frac{1}{n}\mathbf{I} & \cos 2\pi\frac{2}{n}\mathbf{I} & \cdots & \cos 2\pi\frac{n-1}{n}\mathbf{I} \\ \mathbf{0} & \sin 2\pi\frac{1}{n}\mathbf{I} & \sin 2\pi\frac{2}{n}\mathbf{I} & \cdots & \sin 2\pi\frac{n-1}{n}\mathbf{I} \\ \mathbf{I} & \cos 4\pi\frac{1}{n}\mathbf{I} & \cos 4\pi\frac{2}{n}\mathbf{I} & \cdots & \cos 4\pi\frac{n-1}{n}\mathbf{I} \\ \mathbf{0} & \sin 4\pi\frac{1}{n}\mathbf{I} & \sin 4\pi\frac{2}{n}\mathbf{I} & \cdots & \sin 4\pi\frac{n-1}{n}\mathbf{I} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \sin\frac{n-1}{2}2\pi\frac{1}{n}\mathbf{I} & \sin\frac{n-1}{2}2\pi\frac{2}{n}\mathbf{I} & \cdots & \sin\frac{n-1}{2}2\pi\frac{n-1}{n}\mathbf{I} \end{pmatrix}$$

The following theorem is the main result of this paper.

Theorem 3.1. Let $\mathbf{Y} = \{\mathbf{Y}_t, t \in Z\}$ be a T-dimensional stationary real time series such that $E\mathbf{Y}_t = 0$ and $E|Y_t(l)|^2 < \infty, l = 0, \ldots, T-1$, and $\mathbf{R}(t)$ and $\mathbf{h}(w)$ be its autocovariance matrix and spectral matrix of $\mathbf{Y}$, respectively. Then, for large n and under regularity condition, $\Lambda_n(w_j)$ is approximately equal to $2\pi\mathbf{h}(w_j)$, where $\Lambda_n(w_j)$ is the eigen-value matrix of Γ_n, for $w_j = \frac{2\pi j}{n}, j = 0, \ldots, n-1$.

Proof. Let $q_{.,i} = [q_{1,i}\mathbf{I}, q_{2,i}\mathbf{I}, \ldots, q_{n,i}\mathbf{I}]$ be the ith (block) column of $\mathbf{Q}$. We have

$$\begin{aligned} q_{.,i}^{'}\mathbf{I}\Gamma_n^s q_{.,j}\mathbf{I} - q_{.,i}^{'}\mathbf{I}\Gamma_n q_{.,j}\mathbf{I} &= \sum_{m=1}^{M}[\mathbf{R}(m) - \mathbf{R}(n-m)] \\ &\quad \times \sum_{m=1}^{M}[q_{k,i}\mathbf{I}q_{n-m+k,j} + q_{n-m+k,i}\mathbf{I}q_{k,j}] \\ &= \sum_{m=1}^{M} a_m[\mathbf{R}(m) - \mathbf{R}(n-m)], \end{aligned}$$

where $a_m = \sum_{k=1}^{m}[q_{k,i}q_{n-m+k,j} + q_{n-m+k,i}q_{k,j}]$. Now consider the (l, k)th element of the matrix $q_{.,i}^{'}\mathbf{I}\Gamma_n^s q_{.,j}\mathbf{I} - q_{.,i}^{'}\mathbf{I}\Gamma_n q_{.,j}\mathbf{I}$,since $q_{s,i}q_{r,j} \le \frac{2}{n}$ for all s, $i, r, j \in \{1, 2, \ldots, n\}$, then we have

$$\begin{aligned} \left|\left(q_{.,i}^{'}\mathbf{I}\Gamma_n^s q_{.,j}I - q_{.,i}^{'}\mathbf{I}\Gamma_n q_{.,j}\mathbf{I}\right)_{l,k}\right| &= \left|\sum_{m=1}^{M} a_m\left(\mathbf{R}(m) - \mathbf{R}(n-m)\right)_{l,k}\right| \\ &= \left|\sum_{m=1}^{M} a_m E\left[\left(Y_{t+m}(l) - Y_{t+n-m}(l)\right)\left(Y_t(k)\right)\right]\right| \\ &\le \frac{4}{n}\sum_{m=1}^{M} m\left|\gamma_{lk}(m) - \gamma_{lk}(n-m)\right| \\ &\le \frac{4}{n}\left\{\sum_{m=1}^{M} m\left|\gamma_{lk}(m)\right| + \sum_{h=M+1}^{n} M\left|\gamma_{lk}(h)\right|\right\}, \end{aligned} \tag{1}$$

where $\gamma_{lk}(m) = EY_{t+m}(l)Y_t(k)$.

As n increases, the limit of the first term in (3.6) is zero by lemma 3.1.4, in Fuller book, and the limit of the second term is zero by the absolute summability of $\Gamma(h)$. Thus $q_{.,i}^{'}\mathbf{I}\Gamma_n^s q_{.,j}\mathbf{I} - q_{.,i}^{'}\mathbf{I}\Gamma_n q_{.,j}\mathbf{I}$ converges to zero matrix element by element, and so $\mathbf{Q}^{'}\Gamma_n^s\mathbf{Q} - \mathbf{Q}^{'}\Gamma_n\mathbf{Q}$ converges to zero matrix. This means $\mathbf{Q}^{'}\Gamma_n^s\mathbf{Q}$ is approximately equals to $\mathbf{Q}^{'}\Gamma_n\mathbf{Q}$. So, $\mathbf{Q}^{'}\Gamma_n\mathbf{Q} - 2\pi\mathbf{D}_n$ goes to zero using this fact and lemma 3.2. We have proved the theorem.

4 Numerical example

Let $\{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n\}$ be a sample of size n from $\{\mathbf{Y}_t\}$. We assume that $E\mathbf{Y}_t = 0$. Let $n = Mm$, where M and m are integers. Divide the data into M groups, where each group consists of m observations, and let the observation in the l-th group $(l = 1, ..., M)$ be denoted by the $mT \times 1$ vector $\widetilde{\mathbf{Y}}_l$, where

$$\widetilde{\mathbf{Y}}_l = (\mathbf{Y}_{lm}', \mathbf{Y}_{lm-1}', ..., \mathbf{Y}_{(l-1)m+1}')' , \qquad l = 1, ..., M. \tag{2}$$

We estimate the $mT \times mT$ block-Toeplitz covariance matrix Γ_m of order m by

$$\widehat{\Gamma}_m = \frac{1}{M}\sum_{j=1}^{M} \widetilde{\mathbf{Y}}_j \widetilde{\mathbf{Y}}_j'. \tag{3}$$

Let $\widehat{\Lambda}_m(\omega_j)$, $j = 0, ..., m-1$ be the eigenvalue-matrices of $\widehat{\Gamma}_m$ and assume that m is odd. We consider

$$\widehat{\mathbf{h}}_m(\omega_l) = \frac{1}{4\pi}\sum_{j=0}^{m-1} \widehat{\mathbf{A}}_m(\omega_j, \omega_l), \tag{4}$$

as estimators of $\mathbf{h}_m(\omega_l)$, where

$$\widehat{\mathbf{A}}_m(\omega_j, \omega_l) = \frac{2}{m}\sum_{t=1}^{m}\sum_{s=1}^{m} \mathbf{W}_m^{t*}(\omega_j)\widehat{\Lambda}_m(\omega_j)\mathbf{W}_m^{s}(\omega_j)\cos(t-s)\omega_l.$$

(Refer to Nematollahi and Subba Rao (2002), for more details.)

As an example, let $\{\mathbf{Y}_t\}$ be a bivariate stationary series generated from the model

$$\mathbf{Y}_t + \mathbf{A}\mathbf{Y}_{t-1} = \mathbf{e}_t \tag{5}$$

where $\mathbf{A} = \begin{bmatrix} -0.16 & 0.15 \\ -0.14 & -0.15 \end{bmatrix}$ and $\mathbf{e}_t$ is a bivariate Gaussian white noise with mean zero and variance covariance matrix $\Sigma = \begin{bmatrix} 1.19 & 0 \\ 0 & 2.15 \end{bmatrix}$. The spectral density matrix of $\mathbf{Y}_t$ is given by (See, Brockwell and Davis, 1991)

$$\mathbf{h}(\omega) = [\mathbf{I} + \mathbf{A}e^{-i\omega}]^{-1}\Sigma[\mathbf{I} + \mathbf{A}e^{i\omega}]^{'-1}, \ 0 \leq \omega \leq 2\pi. \tag{6}$$

Let $\mathbf{h}(\omega) = [h_{jk}(\omega)]_{j,k=0,1}$. We note $h_{00}(\omega)$ and $h_{11}(\omega)$ are real valued functions and $h_{01}(\omega) = \overline{h_{10}(\omega)}$, is a complex valued function as the cross spectral density function.

First we generated 500 observations $\mathbf{Y}_t, \ t = 1, 2, ..., 500$ from model (4.4) and collected them in 100 groups, with 5 elements in each group, i.e., $M = 100$, $m = 5$. The estimate $\widehat{\mathbf{h}}_m(\omega_l)$ is calculated using the formulae (4.3) with $\omega_j = \frac{2\pi j}{5}$, $j = 0, ..., 4$. The above estimate is computed at the frequencies $\omega_l = l\pi$, $l = 0(0.1)1$. Figure 1 show the logarithm of theoretical density and truncated estimate of $h_{00}(\omega_l)$, respectively. The graph related to $h_{11}(\omega)$ is similar.

References

Brockwell, P. J. and Davis, R. A. (1991). *Time Series : Theory and Method*. Springer Verlag, New York.

Fuller, W. A. (1976) *Introduction to Statistical Time Series* . New York, Wiley.

Hannan, E. J. (1970) *Multiple Time Series*. New York, Wiley.

Hannan, E. J. and Wahlberg, B. (1989) Convergence rates for inverse Toeplitz forms. *J. Multivariate Anal.* **, 31(1),** 127-135.

Nematollahi, A. R. and Subba Rao, T. (2002) On The Spectral Estimation of Periodically Correlated (Cyclostationary) Time Series, Submitted.

Priestley, M. B. (1981) *Spectral Analysis and Time Series*. London, Academic Press

Subba Rao, T. and Gabr, M. M. (1989) The estimation of spectrum, inverse spectrum and inverse autocovariances of a stationary time series. *J. Time Series Anal.,* **10(2)**, 183-202.

Fig. 1.

Nontrivial Application of Jacknife to Problems in Ratio Estimation

Niroumand, H. A.

A11061

Faculty of mathematical sciences, Ferdowsi University of Mashhad, Mashhad, Iran.

Abstract. Let $\hat{\theta_1}$ and $\hat{\theta_2}$ be estimators for θ. Then for any real number $R \neq 1$ we define,

$$G(\hat{\theta_1}, \hat{\theta_2}) = \frac{\hat{\theta_1} - R\hat{\theta_2}}{1 - R}$$

as generalized Jackknife. In considering $G(\hat{\theta_1}, \hat{\theta_2})$ as an estimator for θ when R is known, the question which immediately arises is the manner in which $\hat{\theta_1}$ and $\hat{\theta_2}$ should be selected.

Within the class of estimators for which R is positive and fixed we would desire that $\hat{\theta_1}$ and $\hat{\theta_2}$ have a high positive correlation. On the other hand it would appear

that in the set of all $G(\hat{\theta_1}, \hat{\theta_2})$ one would prefer to have $R < 0$ and $\hat{\theta_1}$ and $\hat{\theta_2}$ negatively correlated.

In this article we shall examine some nontrivial application of Jackknife where additional information is incorporated in the G estimator through the parameter R.

Keywords. Jackknife, Ratio, Estimator.

1 Introduction

In the theory of sampling there is a strong emphasis placed upon the use of auxiliary information. An example of this fact which is of interest here is the use of auxiliary information to improve the precision of estimates through consideration of a population ratio, say ρ. More specification often a situation exists where the ratio of a variable Y to another variable X is believed to have a smaller variance than the Y variable alone. Suppose, for example, one were interested in the value of a population total $T(Y)$. Rather than estimate this total directly from the sample it may be better to estimate $\rho = \frac{T(Y)}{T(X)}$ from the sample and then multiply it by the known total, $T(X)$ to estimate the total $T(Y)$. This is called ratio estimatetion. The ratio estimators are usually biased, though in many situation only negligibly so. On the other hand the bias may be considering in surveys with many strata of small or moderate samples within strata if it is deemed appropriate to use separate ratio estimators for each stratum. When it is considered to be important that proper confidence statements be made it is of necessary that the bias of an estimator be negligibly small. Consequently we must give considerable attention to the development of unbiased or approximately unblased ratio estimators.

In simple random sampling the bias of the ratio estimator r is

$$E(r) - P = -[E(\bar{X})]^{-1} Cov(r, \bar{X})$$

Note that the bias associated with the estimator for the total of Y, $\hat{T}(Y) = T(X)(\frac{\bar{Y}}{\bar{X}})$ is given by $Bias[\hat{T}(Y)] = T(X)Bias(r)$.

Also note that practically speaking r is never unblased since this occurs only if r and $\bar{X}$ are uncorrelated, a situation which seldom occurs in practice.
The decision to use a ratio estimator in hopes of improving the precision is ordinarily based on consideration of the coefficients of variation for the variables X and Y upon the correlation believed to be present between the two. In general the ratio estimators is useful if the characters X and Y have a correlation coefficient which exceeds $1/2$.

After the decision T use a ratio estimators has been made the evaluations of various modifications to the classical estimator which exist will, of necessity, opened upon the assumed model for the relationship between Y and X belongs.
Durbin [1] examined ratio estimators of from $r = \frac{\bar{Y}}{\bar{X}}$, where the regression of Y on X is linerar and X is normally distributed. He considers an application of Quenouille's method which splits the sample into two equal size sets to yield,

$$r_1 = \frac{\bar{Y}_1}{\bar{X}_1} \qquad and \qquad r_2 = \frac{\bar{Y}_2}{\bar{X}_2}$$

where

$$\bar{Y} = \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2) \qquad , \ \bar{X} = \frac{1}{2}(\bar{X}_1 + \bar{X}_2)$$

Then the new estimate, $\hat{\rho}_2$, of $\rho = \frac{E(X)}{E(Y)}$ is

$$\hat{\rho}_2 = 2r - \frac{1}{2}(r_1 + r_2)$$

Suppose

$$\bar{Y} = a + b\bar{X} + U$$

where the $Var(U) = \sigma$, a nonrandom quality of $O(n^{-1})$ and $E[U|\bar{X}] = 0$
Hence, $\rho = \frac{E(X)}{E(Y)} = b + \frac{a}{E(\bar{X})}$ and since $E[U|\bar{X}] = 0$,

$$E(r) = aE(\bar{X}^{-1}) + b \tag{1}$$

Consequently the bias in r is determined by the degree to which $E[\bar{X}^{-1}]$ differs from $(E[\bar{X}])^{-1}$.

1.1 Normal auxiliary

Suppose that $\bar{X}$ is a normal variable with variance h, which is $O(n^{-1})$, and units of measurement R chosen so that $E(\bar{X}) = 1$. Then let $\bar{X} = 1 - \xi$ and hence for sufficiently large n we have

$$E[\bar{X}^{-1}] = E(1 + \xi + \xi^2 + \xi^3 + \ldots)$$

Taking the first four nonvanishing terms we find

$$E[\bar{X}^{-1}] = 1 + h + 3h^2 + 15h^3 + O(n^{-4}) \tag{2}$$

Similarly

$$E[\bar{X}^{-2}] = E(1 + 2\xi + 3\xi^2 + 4\xi^3 + \ldots) = 1 + 3h + 15h^2 + 105h^3 + O(n^{-4}) \tag{3}$$

If we put (1) in (1-1) the bias in r may be determined as

$$E(r) - \rho = aE(\bar{X}^{-1}) + b - (a + b) = a(h + 3h^2 + 15h^3)$$

Neglecting terms of $O(n^{-4})$.
Further, since $Var(\bar{X}_1) = Var(\bar{X}_2) = 2h$, We may replace h by $2h$ in (2) and (3) and obtain

$$E[\bar{X}_i^{-1}] = 1 + 2h + 12h^2 + 120h^3$$

and

$$E[\bar{X}_i^{-2}] = 1 + 6h + 60h^2 + 840h^3 \quad , \quad i = 1, 2$$

Thus if $\hat{\theta}_2 = \frac{r_1 + r - 2}{2}$, Then its bias is given by $E(\hat{\theta}_2) - \rho = a(2h + 12h^2 + 120h^3)$.
Hence if we employ

$$\hat{\theta}_1 = r$$

and

$$\hat{\theta}_2 = \frac{r_1 + r_2}{2}$$

of the from

$$\frac{\hat{\theta}_1 - R\hat{\theta}_2}{1-R}$$

Then to eleminate the to terms of order $O(n^{-4})$ the appropriate choice of R is

$$R = \frac{1+3h+15h^2}{2(1+6h+60h^2)} \tag{4}$$

Selecting $R = \frac{1}{2}$ leads to the estimator $\hat{\rho}_2$ Studied by Durbin [1]. For small h the above expression for R is quite near $\frac{1}{2}$.
Using (3) Durbin (1959) has shown that, not only is the bias of $\hat{\rho}_2$ smaller than of r, but $Var(\hat{\rho}_2) < Var(r)$.
The estimator $\hat{\rho}_3$ combines the same two quantities $\hat{\theta}_1$ and $\hat{\theta}_2$, in the same fashion with a further improvement in the bias let the sample of pairs (Y_i, X_i) $(i = 1, 2, \ldots, n)$ be split at random into N groups each of size M.
Then we get the estimator

$$\hat{\rho}^j = \frac{\bar{Y}^j}{\bar{X}^j}$$

from the sample after omitting the j^{th} group, where

$$\bar{Y}^j = (n\bar{Y} - M\bar{Y}_j)/(n-M)$$

$$\bar{X}^j = (n\bar{X} - M\bar{X}_j)/(n-M)$$

and $\bar{Y}^j$ and $\bar{X}^j$ are the sample means for the j^{th} group. Then Quenouille's [2] estimator is

$$\hat{\rho}_Q = Nr - \frac{N-1}{N}\sum_{1}^{N}\hat{\rho^j} \tag{5}$$

and $Bias(\hat{\rho}_Q)$ and $Var(\hat{\rho}_Q)$ are both decreasing function of N. For $N = 2$, $\hat{\rho}_Q$ Become $\hat{\rho}_2$, The estimator given and student by Durbin [1]. Consequently the indicated optimal choice of $\hat{\theta}_2$ is (corresponding to $N = n$)

$$\hat{\theta}_2 = \frac{1}{n}\sum_{j=1}^{n}\hat{\rho}^j$$

Since the value of h is assumed to be known and h is $O(n^{-1})$, we shall consider the case in which $h = \frac{c}{n}$ for a known constant c. This requires us to choose

$$R = \frac{a[h+3h^2+15h^3]}{a[c/(n-1)+3c^2/(n-1)^2+15c^3/(n-1)^3]}$$

as the proper parameter in the estimator $G(\hat{\theta}_1, \hat{\theta}_2)$ This yields

$$G(\hat{\theta}_1, \hat{\theta}_2) = \rho_4 = \frac{\hat{\theta}_1 - R\hat{\theta}_2}{1-R}$$

Thus ρ_4 would appear to be the best of the estimators which we are considering here.

1.2 Comparison of ρ_3 and ρ_4

The following result will be useful for this comparison,

$$\begin{aligned}
Bias(r) &= a(h+3h^2+15h^3) \stackrel{\triangle}{=} aB(r)\\
Var(r) &= a^2(h+8h^2+69h^3)+\sigma(1+3h+15h^3+105h^4)\\
&\stackrel{\triangle}{=} aS_1(r)+\sigma S_2(r)\\
Bias(\hat{\rho_2}) &= a(6h^2+90h^3) \stackrel{\triangle}{=} aB(\hat{\rho_2})\\
Var(\hat{\rho_2}) &= a^2(h+4h^2+12h^3)+\sigma(1+2h+8h^2+108h^3)\\
&\stackrel{\triangle}{=} a^2S_1(\hat{\rho_2})+\sigma S_2(\hat{\rho_2})
\end{aligned}$$

The estimator $\hat{\rho_2}$ is unbiased to $O(n^{-4})$. In the variance of $\hat{\rho_3}$ let

$$\hat{\rho_3} = cr - d\frac{r_1+r_2}{2}\ ,$$

where $c=\frac{1}{1-R}, d=\frac{R}{1-R}, c-d=1$ (Note that for we have $R=\frac{1}{2}, c=2, d=1$). Using the linear model introduced previousely and splitting the sample as before we have

$$U=\frac{1}{2}(u_1+u_2)\ ,\ \bar{Y_i}=a+hX_i+u_i\ ,\ and\ \ r_i=\frac{\bar{Y_i}}{\bar{X_i}}\ ,\ i=1,2$$

and further that $E(U_i|\bar{X}_i)=0$ and $E(U_i^2|\bar{X}_i)=2\sigma$
Now we may write

$$\begin{aligned}
\hat{\rho_3} &= cb+\frac{c}{\bar{X}}(a+u)-db-\frac{d}{2\bar{X}_1}(a+u_1)-\frac{d}{2\bar{X}_2}(a+u_2)\\
&= b+a\{\frac{c}{\bar{X}}-\frac{d}{2}(\frac{1}{\bar{X}_1}+\frac{1}{\bar{X}_2})\}+\frac{cU}{\bar{X}}-\frac{d}{2}(\frac{U_1}{\bar{X}_1}+\frac{U_2}{\bar{X}_2})
\end{aligned}$$

Hence

$$E(\hat{\rho_3}-b)=aE[\frac{c}{\bar{X}}-\frac{d}{2}(\frac{1}{\bar{X}_1}+\frac{1}{\bar{X}_2})], \tag{6}$$

and when R is known

$$c=\frac{2(1+6h+60h^2)}{1+9h+105h^2}\ \ d=\frac{2(1+3h+15h^2)}{1+9h+105h^2}$$

and therefore $E(\hat{\rho_3}-b)=a+O(n^{-4})$. Next consider

$$E\{\frac{c}{\bar{X}}-\frac{d}{2}(\frac{1}{\bar{X}_1}+\frac{1}{\bar{X}_2})\}^2 \tag{7}$$

$$=E\{\frac{c^2}{\bar{X}^2}-\frac{d^2}{4}(\frac{1}{\bar{X}_1^2}+\frac{1}{\bar{X}_2^2})+(\frac{d^2}{2})(\frac{1}{\bar{X}_1\bar{X}_2})\} \tag{8}$$

$$=c^2(1+3h+15h^2+105h^3)+\frac{d^2}{2}(1+6h+60h^2+840h^3) \tag{9}$$

$$+(\frac{d^2}{2}-2cd)(1+2h+12h^2+120h^3)^2, \tag{10}$$

and

$$E_{\bar{X}}E_{U|\bar{X}}[\{\frac{c(U_1+U_2)}{2\bar{X}}-\frac{d}{2}(\frac{1}{\bar{X}_1}+\frac{1}{\bar{X}_2})\}^2|\bar{X}_1\bar{X}_2] \tag{11}$$

$$= E_{\bar{X}}2\sigma[(\frac{c}{2\bar{x}}-\frac{d}{2\bar{X}})^2+(\frac{c}{2\bar{X}}-\frac{d}{2\bar{X}_2})^2] \tag{12}$$

$$= E[\frac{2c^2}{\bar{x}^2}+d^2\frac{1}{\bar{X}_1^2}+\frac{1}{\bar{X}_2^2})-4cd(\frac{1}{\bar{X}_1\bar{X}_2})] \tag{13}$$

$$= \sigma[c^2(1+3h+15h^2+105h^3)+d^2(1+6h+60h^2+840h^3) \tag{14}$$

$$- 2cd(1+2h=12h^2+120h^3)] \ , \tag{15}$$

Hence, substituting approximate expressions for C^2, D^2, and cd, we obtain

$$E(\hat{\rho}_3-b)=a^2[\frac{1+26h+435h^2+409h^3}{1+18h+291h^2+1890h^3}]+\sigma[\frac{1+28h+471h^2+4680h^3}{1+18h+291h^2+1890h^3}]$$

Consequently

$$Var(\hat{\rho}_3)=Var(\hat{\rho}_3-b)=a^2[\frac{1+8h+117h^2+2046h^3}{1+18h+291h^2+1890h^3}]+\sigma[\frac{1+28h+471h^2+4680h^3}{1+18h+291h^2+189h^3}].$$

The values of the coefficients, S_1, S_2 of a^2 and σ respectively, have been tabulated for several values of h. The coefficients of a in the expression for the bias are also given here in table (1). On bias was calculated for ρ_3 since all terms containing the fourth power and higher in h have been neglected in the orginal approximation and ρ_3 is corrected for bias to this degree. Because of the entries in the lower half of the table are subject to considerable error.

For instance the error in $B(r)$ for $h = 0.050$ is greather than 6. In spite of this, the large values of h have been included for two reasons. First of all these larger values are not unusual in ordinary practical applications. Second, note from table (1) that even at $h = 0.1$ the bias in $\hat{\rho}_2$ is greather than that of r; a disturbing result since $\hat{\rho}_2$ was proposed from a bias reduction standpoint. However, recalling that the approximation made at the outset valid only for small h, this later observation is a commentary on the range of validity of these approxiamtions rather than the Jackknife method.

Table (1): Variance comparisons for normal model

	r			ρ_2			ρ_3		
h	B	S_1	S_2	B	S_1	S_2	B	S_1	S_2
0.01	0.010	0.011	1.032	0.001	0.010	1.021	0.0	0.077	1.100
0.05	0.058	0.079	1.201	0.016	0.061	1.133	0.0	0.331	1.454
0.10	0.145	0.249	1.555	0.150	0.152	1.388	0.0	0.528	1.736
0.15	0.268	0.563	2.142	0.439	0.280	1.844	0.0	0.646	1.900
0.20	0.440	1.072	3.040	0.960	0.456	2.584	0.0	0.722	2.005
0.25	0.672	1.828	4.328	1.781	0.687	3.687	0.0	0.776	2.078
0.30	0.945	2.883	6.085	2.97	0.984	5.236	0.0	0.815	2.130
0.40	1.84	6.096	11.320	6.72	1.808	9.992	0.0	0.870	2.203
0.50	3.125	11.125	19.375	12.75	3.000	12.500	0.0	0.906	2.250
0.75	8.765	34.359	55.584	41.34	8.062	52.562	0.0	0.958	2.318
1.000	19.0	78.0	124.0	96.0	17.0	119.0	0.0	0.987	2.355

bIaS(0)=a.b(0) Variance(0)=$a^2S_1(0) + \sigma S_2(0)$

Conclusions

In each expression the terms of the fourth and higher power in h have been neglected. The breakdown of the necessary approximation for the normal model for the realistically large coefficients of variation is inducement to examine a different model. Furthermore, since practically all auxiliary random variables, X, which are used in real problems are positive, the normal model is not realistic when h is near.

References

[1] Durbin, j. (1959),*a note on the application of Quenouille's method of bias reduction to the estimation of ratios*, biOmetrika 64, 477-480.

[2] Quenouille, M. (1956), Note on bias in estimation, biometrika, 43.

[3] RaO, j.N.K. (1965),*a note the estimation of ratios by Quenouille method*, biometrika, 52.

On the Convergence of the Objective Function of a Sample Size Determination Problem

Pezeshk, H.[1] **and Gittins, J.**[2]

P11002

[1] Department of Statistics, University of Tehran, Iran.
[2] Department of Statistics, University of Oxford, UK.

Abstract. An important question in the planning of trials is how large to make the trial. The problem may be formulated formally in statistical terms. There have been a number of papers, from both the frequentist and Bayesian points of view, on this subject (listed, for example, by Adcock (1997)).

Several authors have recognized the value of using prior distributions rather than point estimates in sample size calculations. Bayesian methods, which use a prior distribution for the unknown parameters, may be divided into two groups of procedures: methods which are inferential (see, for example, Joseph *et al* (1997)), and fully Bayesian (or decision theoretic) methods which treat the problem as a decision problem and employ a loss or utility function (see, for example, Lindley (1997), Pezeshk *et al* (2001), and Gittins and Pezeshk (2000)).

In this paper we discuss the convergence of the objective function and hence also the convergence of the size of a trial for which the data are assumed to come from a normal distribution for which the mean and the variance are both unknown. The objective function is the expected benefit of conducting the trial under consideration minus the cost of it.

Keywords. Sample Size Determination, Fully Baysian Approach, Normal Distribution, Expected Net Benefit.

1 Introduction

An important question in the planning of trials is how large to make the trial. The problem may be formulated formally in statistical terms. There have been a number of papers, from both the frequentist and Bayesian points of view, on this subject (listed, for example, by Adcock (1997)).

Several authors have recognized the value of using prior distributions rather than point estimates in sample size calculations. Bayesian methods, which use a prior distribution for the unknown parameters, may be divided into two groups of procedures: methods which are inferential (see, for example, Joseph *et al* (1997), Spiegelhalter and Freedman (1986)), and fully Bayesian (or decision theoretic) methods which treat the problem as a decision problem and employ a loss or utility function (see, for example, Lindley (1997), Stallard (1998), Pezeshk and Gittins (1999), and Gittins and Pezeshk (2000b)).

In section 2 we state the sample size problem and introduce the notation. In section 3 we establish our objective function, which is the expected net benefit of conducting the trial. Section 4 shows the consistency of the sample size calculation in the

unknown variance case by illustrating the convergence of the reward function for a sequence of unknown variance cases to the reward function for the limiting case in which σ^2 is known. These calculations also show that convergence to the known variance case is not particularly fast, thus demonstrating the practical importance of the unknown variance methodology.

2 The Sample Size Problem

Consider paired observations (X_1, Y_1), (X_2, Y_2), ... and assume that $Z_i = X_i - Y_i$ $(i = 1, 2, ...)$ has a normal density which is $N(\delta, \sigma^2)$, where both δ and σ^2 are unknown. (Note that the assumption of paired observations is for illustration only. The methodology applies equally well with an unpaired observations.)

Let m be the number of subsequent users of the new service or the new treatment for which the trial is being carried out. This depends on the posterior distribution of δ and σ^2 as discussed at the end of this section.

For every user who goes on to use the new service as a result of the trial there is a benefit. The objective (or expected net benefit) function $r(n)$, for a trial with n pairs of users, is the total expected benefit from the resulting change in the number of users of the new service minus the cost of the trial. The benefit per user may simply be a constant b independent of δ. The question is, how many observations may maximize the total expected benefit?

Let the cost of carrying out a trial with n pairs of observations be $cn + d$ if $n > 0$, and 0 if $n = 0$. Thus d is a set-up cost. We can proceed to calculate the optimal trial size n^* as though $d = 0$. The trial should be carried out only if $r(n^*) > 0$, where $r(n^*)$ now includes the set-up cost. To simplify our discussion we shall assume from now on that $d = 0$.

Following O'Hagan (1994), let us assume that the prior density functions for δ and σ^2 are of the form shown below

$$\pi(\delta) = \frac{1}{B(1/2, g/2)}(aw)^{-1/2}\{1 + \frac{(\delta - \mu)^2}{aw}\}^{-\frac{g+1}{2}},$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$, and

$$\pi(\sigma^2) = (\frac{a}{2})^{\frac{g}{2}} \frac{1}{\Gamma(g/2)}(\sigma^2)^{-\frac{g+2}{2}} exp(-\frac{a}{2\sigma^2}). \quad (1)$$

The latter is a kind of inverse chi-squared distribution, because the distribution of a/σ^2 is chi-squared with g degrees of freedom.

Applying Bayes' theorem the posterior densities turn out to be

$$\pi^n(\delta|z) = (w'a')^{-1/2} \frac{1}{B(1/2, g'/2)}\{1 + (\delta - \mu')^2/w'a'\}^{-\frac{g'+1}{2}}.$$

and

$$\pi^n(\sigma^2|z) = \frac{(\frac{a'}{2})^{g'/2}}{\Gamma(g'/2)}(\sigma^2)^{-\frac{g'+2}{2}} exp(-\frac{a'}{2\sigma^2}), \quad (2)$$

where $z_i = x_i - y_i$ $(i = 1, ..., n)$, z is the vector of observations, $\overline{z}_n = \sum_i z_i/n$, $s^2 = \sum_i (z_i - \overline{z}_n)^2$, and

$$w' = \frac{w}{1+nw}, \qquad \mu' = \frac{\mu + nw\overline{z}_n}{1+nw}$$

$$g' = g + n, \qquad a' = a + s^2 + \frac{n(\overline{z}_n - \mu)^2}{1+nw}.$$

Therefore $(\delta - \mu')/(w'a'/g')^{-1/2}$ has a Student t-distribution with g' degrees of freedom. So the posterior mean and variance of δ (provided $g' > 2$) are

$$\mu'(\overline{z}_n) = \frac{\mu + nw\overline{z}_n}{1+nw}, \qquad \tau'^2(\overline{z}_n, s^2) = \frac{w'a'}{g'-2}.$$

Let us suppose that m, the number of subsequent users, depends on the mean μ' and the standard deviation τ' of the posterior distribution for δ as shown below.

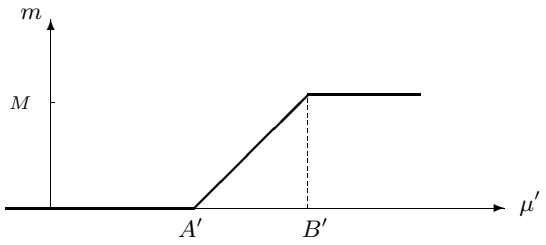

Figure 1: Number of subsequent users,
where $A' = A + 1.5\tau'(\overline{z}_n, s^2)$ and $B' = B + 1.5\tau'(\overline{z}_n, s^2)$.

Here M is the expected total number of users, given a substantial improvement in performance. A and B are two parameters which must be estimated. Their values depend on the difference between the expected cost of the new service and that of the current service.

This function corresponds to assuming that each individual has a personal threshold difference between A and B, and is prepared to switch to the new service or the new treatment provided that the apparent difference between the two services exceeds this threshold by at least 1.5 standard deviations of the posterior distribution for the difference.

Using some algebra, we see that

$$f(\overline{z}_n, s^2) = \frac{\Gamma((g+n)/2)}{\Gamma(g/2)\Gamma(1/2)}\left(\frac{a^{g/2}}{\Gamma(\frac{n-1}{2})}\right)\left(\frac{n}{1+nw}\right)^{\frac{1}{2}} \times (s^2)^{\frac{n-1}{2}-1}\left(a + s^2 + \frac{n(\overline{z}_n - \mu)^2}{1+nw}\right)^{-\frac{g+n}{2}}. \quad (3)$$

Let the number of subsequent users of the new service be of the form shown in figure 1, so that

$$m = \begin{cases} 0 & \mu' < A' \\ \frac{M}{B'-A'}(\mu' - A') & A' < \mu' < B' \\ M & B' < \mu'. \end{cases}$$

Pezeshk *et al* (2001) showed that the objective (or expected net benefit) function is

$$\begin{aligned} R(n) = {} & \frac{1}{(B-A)}\frac{\Gamma((g+n)/2)}{\Gamma(g/2)\Gamma(1/2)}\left(\frac{a^{g/2}}{\Gamma(\frac{n-1}{2})}\right)\left(\frac{n}{1+nw}\right)^{\frac{1}{2}}\int_0^{\infty}\int_{H_1}^{H_2}(\mu' \\ & - A - 1.5\tau')(s^2)^{\frac{n-1}{2}-1} \times \left(a + s^2 + \frac{n(\overline{z}_n - \mu)^2}{1+nw}\right)^{-\frac{g+n}{2}} d\overline{z}_n ds^2 \\ & + \frac{\Gamma((g+n)/2)}{\Gamma(g/2)\Gamma(1/2)}\left(\frac{a^{g/2}}{\Gamma(\frac{n-1}{2})}\right)\left(\frac{n}{1+nw}\right)^{\frac{1}{2}}\int_0^{\infty}\int_{H_2}^{\infty}(s^2)^{\frac{n-1}{2}-1} \\ & \times \left(a + s^2 + \frac{n(\overline{z}_n - \mu)^2}{1+nw}\right)^{-\frac{g+n}{2}} d\overline{z}_n ds^2 \\ & - Cn. \end{aligned} \quad (4)$$

where $R(n) = r(n)/(Mb)$, $C = c/(Mb)$, and H_1 and H_2 are the values of $\overline{z}_n$ for which, respectively,

$$\overline{z}_n = \frac{(A + 1.5(\frac{w'a'}{g'-2})^{\frac{1}{2}})(1+nw) - \mu}{nw},$$

and

$$\overline{z}_n = \frac{(B + 1.5(\frac{w'a'}{g'-2})^{\frac{1}{2}})(1+nw) - \mu}{nw}.$$

Note that a' depends on $\overline{z}_n$ and s^2. Solving for $\overline{z}_n$ in the first equation produces a quadratic equation with two real roots, the larger of which may be shown to be H_1. H_2 may be found in similar fashion. Note also that for sufficiently small values of n the set of values of $\overline{z}_n$ for which $\mu' > A + 1.5\tau'$ is the interval bounded by the two roots of the quadratic.

Figure 2 illustrates the variation of $R(n)$ as a function of n. The optimal sample size n^* for this case is 48.

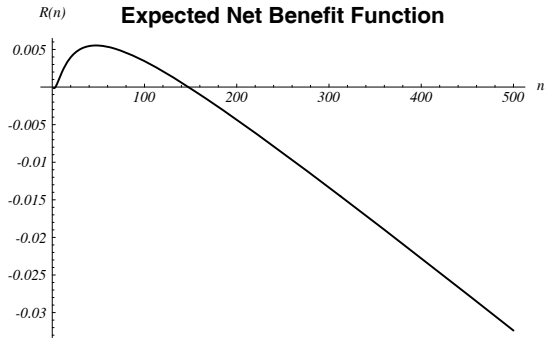

Fig. 1. Expected net benefit when $a = w = 1$, $\mu = 1$, $g = 5$, $c/(Mb) = 0.0001$, and $(A, B) = (2, 2.5)$; commercial benefit function.

3 Convergence to the Known Variance Case

Gittins and Pezeshk (2000b) show that the objective function for normally distributed data with known variance, is

$$\begin{aligned} R(n) = & \frac{1}{B-A}\int_{h_1(A,n)}^{h_2(B,n)} \{\mu + n^{\frac{1}{2}}\tau(\sigma^2/\tau^2+n)^{-1/2}u\}(\frac{1}{2\pi})^{\frac{1}{2}}e^{-\frac{1}{2}u^2}du \\ & - \frac{A}{B-A}\int_{h_1(A,n)}^{h_2(B,n)} (\frac{1}{2\pi})^{\frac{1}{2}}e^{-\frac{1}{2}u^2}du \\ & - \frac{1}{B-A}1.5(\sigma^2/\tau^2+n)^{-1/2}\int_{h_1(A,n)}^{h_2(B,n)} (\frac{1}{2\pi})^{\frac{1}{2}}e^{-\frac{1}{2}u^2}du \\ & + \int_{h_2(B,n)}^{\infty} (\frac{1}{2\pi})^{\frac{1}{2}}e^{-\frac{1}{2}u^2}du - Cn, \end{aligned}$$

where

$$h_1(A,n) = \frac{\{A + 1.5\sigma(\sigma^2/\tau^2+n)^{-1/2} - \mu\}(\sigma^2/\tau^2+n)^{1/2}}{\tau n^{\frac{1}{2}}},$$

$$h_2(B,n) = \frac{\{B + 1.5\sigma(\sigma^2/\tau^2+n)^{-1/2} - \mu\}(\sigma^2/\tau^2+n)^{1/2}}{\tau n^{\frac{1}{2}}}.$$

Here we show that for a sequence of unknown variance cases for which the mean and variance of the prior distribution for δ and the mean of the prior distribution for σ^2 are held fixed, the variance of the prior distribution for σ^2 is decreasing, and the prior distribution for δ tends to normal, n^* is decreasing and tends to the value for the known σ^2 limit.

Recall that the prior density function for δ is such that

$$\frac{\delta - \mu}{\sqrt{\frac{w\alpha}{g}}} \sim t_g, \tag{5}$$

which converges in distribution to $N(0,1)$ as $g \to \infty$.

From (3) one can derive the (predictive) density function for $\overline{Z}_n$. This is

$$f(\overline{z}_n) = \frac{1}{B(1/2, g/2)} \Big(\frac{a(1+nw)}{n}\Big)^{-1/2} [1 + \frac{(\overline{z}_n - \mu)^2}{\frac{a(1+nw)}{n}}]^{-\frac{g+1}{2}}. \tag{6}$$

Putting $T = \frac{\overline{Z}_n - \mu}{\sqrt{\frac{a(1+nw)}{ng}}}$ it follows that the density function for T is

$$\frac{1}{B(1/2, g/2)} \ \frac{1}{\sqrt{g}} \ [1 + \frac{t^2}{g}]^{-\frac{g+1}{2}}, \tag{7}$$

so that $T \sim t_g$.

Now note that it follows from equation (1) that the mean and variance of the prior distribution for σ^2 are $a/(g-2)$ and $2a^2/[(g-2)^2(g-4)]$.

With a tight prior distribution for σ^2 we are close to the situation where σ^2 is known to be equal to the mean of this distribution. Thus the optimal sample size calculated with a tight prior distribution for σ^2 should result in a value close to the one for the known variance case. We can use this fact to check the consistency of our sample size calculations for the case where the variance of the observations is unknown.

Let a and g tend to infinity, subject to $\frac{a}{g-2} = \rho^2$ and $\frac{wa}{g-2} = \tau^2$, where ρ and τ are fixed. The following observations up to and including the theorem are all based on this assumption.

We have

$$T \xrightarrow{D} N(0,\ 1), \qquad \overline{Z}_n \xrightarrow{D} N(\mu,\ \tau^2 + \frac{\rho^2}{n}), \qquad \sigma^2 \xrightarrow{P} \rho^2. \tag{8}$$

It follows that the predictive distribution of $\overline{Z}_n$ in the unknown variance case tends to the one for the known variance case.

Since both the prior distribution of δ and the predictive distribution of $\overline{Z}_n$ in the unknown variance case tend to those for the known variance case, it is not difficult to show that $m\mu'$ converges in distribution to its distribution for the known variance case and the expected net benefit function tends to the one for known variance. It follows that the optimal sample size tends to its value for the known σ^2 limit. We shall prove this through the following lemmas and theorem. We use the subscript u to indicate that the parameter belongs to the case of unknown variance and the subscript k to indicate the known variance case.

Lemma: The posterior mean of δ for unknown variance converges in distribution to its value for known variance. In symbols

$$\frac{\mu + nw\overline{Z}_u}{1 + nw} \ \xrightarrow{D} \ \frac{\rho^2\mu + n\tau^2\overline{Z}_k}{\rho^2 + n\tau^2}. \tag{9}$$

Proof: From (8) we have $\overline{Z}_u \xrightarrow{D} \overline{Z}_k$. It follows that

$$\frac{\frac{a}{g-2}\,\mu + n\,\frac{wa}{g-2}\,\overline{Z}_u}{\frac{a}{g-2} + n\,\frac{wa}{g-2}} \xrightarrow{D} \frac{\rho^2\mu + n\tau^2\overline{Z}_k}{\rho^2 + n\tau^2}.$$

Thus

$$\frac{\mu + nw\overline{Z}_u}{1+nw} \xrightarrow{D} \frac{\rho^2\mu + n\tau^2\overline{Z}_k}{\rho^2 + n\tau^2}, \tag{10}$$

and (9) follows.

Lemma: The posterior variance of δ for the unknown variance case converges in probability to its value for known variance. In symbols

$$\frac{\frac{w}{1+nw}(a + s^2 + \frac{n(\overline{Z}_u-\mu)^2}{1+nw})}{n+g-2} \xrightarrow{P} \frac{\rho^2\tau^2}{\rho^2+n\tau^2}. \tag{11}$$

Proof: First note that as g tends to infinity $\overline{Z}_u \xrightarrow{P} \mu$. So $(\overline{Z}_u - \mu)^2 \xrightarrow{P} 0$. Also note that $\frac{s^2}{n-1} \xrightarrow{P} \rho^2$ as $g \longrightarrow \infty$. So

$$a + s^2 + \frac{n(\overline{Z}_u - \mu)^2}{1+nw} \xrightarrow{P} a + (n-1)\rho^2. \tag{12}$$

Now note that

$$\frac{w(a+(n-1)\rho^2)}{(1+nw)(n+g-2)} \longrightarrow \frac{wa/(g-2)}{1+nw} = \frac{\rho^2\tau^2}{\rho^2+n\tau^2}, \tag{13}$$

and the lemma follows.

Now we are in a position to state the following theorem.

Theorem: If a and g tend to infinity, subject to $\frac{a}{g-2} = \rho^2$ and $\frac{wa}{g-2} = \tau^2$, where ρ^2 and τ^2 are fixed, then $E(m_u) \longrightarrow E(m_k)$.

Proof: The number of subsequent users of the new treatment, m, is a bounded continuous function of μ'_u and τ'_u, and we have $\mu'_u \xrightarrow{D} \mu'_k$ and $\tau'_u \xrightarrow{P} \tau'_k$ as a and g tend to infinity. The results follow, using the continuous mapping theorem (see,

for example, theorem 2.3, of Durret (1996), p.87).

The above theorem states that the reward functions for unknown variance converge to the corresponding reward functions for known variance.

As an example, let us consider the objective function, with known variance and with the following parameter values.

$$\sigma^2 = 4, \quad \tau = 1.045, \quad \mu = 2.09, \quad \frac{c}{Mb} = \frac{4000}{5000000}, \quad (A, B) = (1.67, 2.51).$$

Calculations show that the optimal sample size is $n^* = 51.53$ and that $r(n^*) = 1.897 \times 10^6$. (Of course n is in practice restricted to integer values, but the convergence of the calculated values is clearer if we relax that condition.)

A sequence of sets of parameter values for unknown variance which converge to this case is shown below.

1)

$$g = 5, \quad a = 12, \quad w = 0.273, \quad \mu = 2.09, \quad (A, B) = (1.67, 2.51);$$

$$\boxed{n^* = 88.31, \qquad r(n^*) = 1.595 \times 10^6,}$$

2)

$$g = 12, \quad a = 40, \quad w = 0.273, \quad \mu = 2.09, \quad (A, B) = (1.67, 2.51);$$

$$\boxed{n^* = 79.23, \qquad r(n^*) = 1.639 \times 10^6,}$$

3)

$$g = 32, \quad a = 120, \quad w = 0.273, \quad \mu = 2.09, \quad (A, B) = (1.67, 2.51);$$

$$\boxed{n^* = 75.78, \qquad r(n^*) = 1.674 \times 10^6.}$$

4)

$$g = 102, \quad a = 400, \quad w = 0.273, \quad \mu = 2.09, \quad (A, B) = (1.67, 2.51);$$

$$\boxed{n^* = 61.31, \qquad r(n^*) = 1.751 \times 10^6.}$$

5)

$$g = 127, \quad a = 500, \quad w = 0.273, \quad \mu = 2.09, \quad (A, B) = (1.67, 2.51);$$

$$\boxed{n^* = 53.45, \qquad r(n^*) = 1.882 \times 10^6.}$$

As expected n^* and $r(n^*)$ approach their value for known σ^2 as a and g tend to infinity.

As an another example, let us consider the objective function, with known variance and with the following parameter values.

$$\sigma^2 = 18.66, \quad \tau = 0.55, \quad \mu = 1.1, \quad \frac{c}{Mb} = \frac{600}{25000000}, \quad (A, B) = (0.9, 1.3).$$

Calculations show that the optimal sample size is $n^* = 956.71$ and that $r(n^*) = 10.960 \times 10^6$.

A sequence of sets of parameter values for unknown variance which converge to this case is shown in table 1.

Table 1: Parameter values for the corresponding unknown variance case

a	g	n^*	$r(n^*)$
56	5	2159.62	8.995×10^6
223.92	12	1859.00	9.134×10^6
373.20	20	1635.76	9.545×10^6

$w = 0.016$, $\mu = 1.1$, and $(A, B) = (0.9, 1.3)$.

As before, here n^* decreases to the known variance value as a and g tend to infinity.

References

1. Adcock, C.J. (1997). Sample Size Determination: A Review. *Statistician* **46,** 261–283.
2. Durrett, R. (1996). *Probability: Theory and Examples, 2nd edn.* Duxbury Press Belmont.
3. Gittins, J.C. & H. Pezeshk, H. (2000a). A Behavioural Bayes Method for Determining the Size of A Clinical Trial. *Drug Information Journal* **34,** pp. 355–363.
4. Gittins, J.C. & H. Pezeshk, H. (2000b). How Large Should A Clinical Trial Be? *The Statistician* **49,** part 2, 177–187.
5. Joseph, L et al. (1997). Bayesian and Mixed Bayesian/Likelihood Criteria for Sample Size Determination. *Statistics in Medicine* **16,** 769–781.
6. Lindley, D.V. (1997). The Choice of Sample Size. *The Statistician* **46,** 129–138.
7. O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics Volume 2B Bayesian Statistics.* Edward Arnold.
8. Pezeshk, H. & Gittins, J.C. (1999) Sample Size Determination in Clinical Trials. *Student* **3,** No.1, 19–26.
9. Pezeshk, H., Kikuchi, T. & Gittins, J.C. (2001) A Decision Theoretic Approach to Sample Size Queston in Clinical Trials. 22^{nd} *Annual Conference The International Society for Clinical Biostatistics, ISCB Stockholm, Sweden* p:1, pp.128.
10. Spiegelhalter, D.J. & Freedman, L.S. (1986). A Predictive Approach to Selecting the Size of a Clinical Trial, Based on Subjective Clinical Opinion. *Statistics in Medicine* **5,** 1–13.
11. Stallard, N. (1998). Sample Size Determination for phase II Clinical Trials based on Bayesian Decision Theory. *Biometrics* **54,** 279–294.
12. Wolfram, S. (1991). *Mathematica: a system for doing mathematics by computer.* Addison Wesley, Redwood City.

Distance Sampling: Line Transect

Salehi, M. M.

A11259

School of Mathematical Science, Isfahan University of Technology, Isfahan, Iran.

Abstract. Distance sampling is a widely-used group of closely related methods for estimating the density and/or abundance of biological populations. The main methods are line transects and point transects. These have been used successfully in trees, shrubs and herbs, insects, amphibians, reptiles, birds, fish, marine and land mammals etc. In both cases, the basic idea is the same. The observer(s) perform a standardized survey along a series of lines or points, searching for objects of interest. For each object detected, they record the distance from the line or point to the object. Not all the objects that the observers pass will be detected, but a fundamental assumption of the basic methods is that all objects that are actually on the line or point are detected. In this article, we focus on line transect.

Keywords. Adaptive Sampling, Detectability, Key Function.

Line transect sampling is described as follows. A line of length L is chosen at random in an area of size A containing N objects. (How would you do this!). An observer(B) moves down the line and looks out for any objects out to distance w from the line on either side and records(estimates) the perpendicular distance y out to each object (C) observed. The aim is to estimate $D = N/A$, or N.

The line transect method was introduced in the 1930's using the radial distance r, but little happened until 1968 when Eberhardt (1968) and Gates et. al. (1968) developed some rigorous models. From 1976 onwards there was a whole string of new developments culminating in the first full length book on the subject "Distance sampling : estimating abundance of biological populations" by Buckland, Anderson, Burnham and Laake (1993), published by Chapman & Hall. The second edition of this book is now out (2001). There is also a computer package available, called **Distance**.

1 Theory

Suppose that each object has the same probability P of being detected from the random transect. (We are not assuming independence of the objects). Let n be the number of objects detected. Then

$$n = \sum_{i=1}^{N} I_i \ ,$$

where $I_i = 1$, with probability P, if object i is detected, and 0 otherwise. Since $E[I_i] = P$,

$$E[n] = \sum_{i=1}^{N} P = NP, \quad \text{and} \quad N = \frac{E[n]}{P} \ . \tag{1}$$

(Note that the I_i do not need to be independent). If an estimate $\hat{P}$ of P is available, then $\hat{N} = n/\hat{P}$.

Now

$$\begin{aligned} P &= \text{pr(object detected} \mid \text{in observation area)pr(in observation area)} \\ &= P_W P_L \text{ , say.} \end{aligned}$$

For any individual object, placing a line at random and seeing if the object is in the observation area, a rectangle $2w \times L$, is equivalent to fixing the line and dropping the object at random. (Again we are not implying that objects are independent). Then the probability that the object falls in the rectangle is

$$P_L = \frac{2Lw}{A} \ .$$

To find P_w we first introduce a key concept called the "detection" function, $g(y)$. We show below that, for $0 \leq y \leq w$, $g(y) = \text{pr(detected} \mid \text{object at distance } y)$. Note that $0 \leq g(y) \leq 1$. We now consider objects in the $L \times 2w$ rectangle. For any such object, let $X = 1$ if it is detected, and $X = 0$ if it is not detected. Let $f_{X,Y}(x, y)$ be the joint distribution of X and Y. (Note that X is discrete but Y is continuous, with $0 \leq y \leq w$). The next step is to prove that

$$g(y) = f_{X|Y}(1|y) \ .$$

Now

$$\begin{aligned} g(y) &= \lim_{h \to 0} \text{pr}(X = 1 | \text{distance in } (y-b, y+b)) \\ &= \lim_{h \to 0} \frac{\text{pr}(X = 1 \text{ and } y - h \leq Y \leq y + h)}{\text{pr}(y - h \leq Y \leq y + h)} \\ &= \lim_{h \to \infty} \left\{ \frac{\int_{y-h}^{y+h} f_{X,Y}(1,t)dt}{\int_{y-h}^{y+h} f_Y(t)dt} \right\} \\ &= \lim_{h \to \infty} \left\{ \frac{f_{X,Y}(1,y)2h}{f_Y(y)2h} + O(h) \right\} \\ &= \frac{f_{X,Y}(1,y)}{f_Y(y)} = f_{X|Y}(1|y) \ . \end{aligned}$$

Since the transect is randomly located,

$$f_Y(y) = \frac{1}{w}, \quad 0 \leq y \leq w.$$

Now

$$\begin{aligned} f_X(x) &= \int_0^w f_{X,Y}(x,y)dy \\ &= \int_0^w f_{X|Y}(x|y) f_Y(y)dy \\ &= \frac{1}{w} \int_0^w f_{X|Y}(x|y)dy \end{aligned}$$

and

$$P_w = f_X(1) = \frac{1}{w}\int_0^w f_{X|Y}(1|y)dy = \frac{1}{w}\int_0^w g(y)dy\ .$$

Hence

$$P = P_w P_L = \frac{2Lw}{A}\cdot\frac{1}{w}\int_0^w g(y)dy = \frac{2L}{A}\lambda\ , \tag{2}$$

where $\lambda = \int_0^w g(y)dy$. We shall also be interested in

$$\begin{aligned} f(y) &= f_{Y|X}(y|1) \\ &= \frac{f_{X,Y}(1,y)}{f_X(1)} \\ &= \frac{f_{X|Y}(1|y)f_Y(y)}{f_X(1)} \\ &= \frac{g(y)}{wP_w} = \frac{g(y)}{\lambda}\ . \end{aligned}$$

This is the density function for Y given the object is detected. We see that $f(y) = g(y)/\lambda$ has the same shape as $g(y)$; the former can be obtained simply by rescaling $g(y)$ so as to integrate to 1.

[Note that $\lambda = wP_w = \int_0^w g(y)dy$, so we can let $w \to \infty$; λ is still well defined as $\int_0^\infty g(y)dy$. Truncation to w is often done afterwards to remove possible outliers.]

Using (1) and (2) gives us

$$D = \frac{N}{A} = \frac{E[n]}{PA} = \frac{E[n]}{2L\lambda}\ . \tag{3}$$

To estimate D we need to estimate λ. *Assumption* : $g(0) = 1$ If $g(0) = 1$, $\lambda = \frac{1}{f(0)}$ and

$$\begin{aligned} \hat{D} &= \frac{n\hat{f}(0)}{2L} \\ &= \frac{n}{2L\hat{\lambda}}\ . \end{aligned} \tag{4}$$

In practice $f(0)$ (and λ) are estimated conditionally on n, so that $\hat{f}(0)$ is a function of $y_1, y_2, \cdots, y_n$. This usually means that $E[\hat{f}(0)|n] \approx E[\hat{f}(0)]$ and, by the following lemma, $cov[\hat{f}(0), n] \approx 0$.

Lemma : If $E[X|Y] = E[X]$, then $cov[X, Y] = 0$.

Proof :

$$\begin{aligned} cov[X,Y] &= E[XY] - \mu_X\mu_Y \\ &= E_Y\{E[XY|Y]\} - \mu_X\mu_Y \\ &= E_Y\{YE[X|Y]\} - \mu_X\mu_Y \\ &= E_Y\{YE[X]\} - \mu_X\mu_Y \\ &= E[Y]E[X] - \mu_X\mu_Y = 0\ . \end{aligned}$$

If $E[\hat{f}(0)] \approx f(0)$ we can use our short-cut delta method= for finding an approximate expression for $var[\hat{D}]$, namely

$$\frac{var[\hat{D}]}{D^2} \approx \frac{var[\hat{f}(0)]}{\{f(0)\}^2} + \frac{var[n]}{\{E[n]\}^2}\ . \tag{5}$$

We need further assumptions to find the above expressions (see later). *Assumptions*

(a) D is constant (More generally, if D is a spatial stochastic Process, we must have $E[D]$ is constant). We don't need the assumption of a random distribution.
(b) The detection function, $g(y)$, is the same for all objects.
(c) Objects are detected at their initial location (i.e. no movement away from the observer if the objects are animals)
(d) Measurements are exact.

Also, the above theory assumes $g(0) = 1$. In the above assumptions, (b) is not so critical as we are only interested in $f(0)$. As far as (b) is concerned, we can allow for the size of the object using $g(y, s)$, where s is the size. For (d), the data can be grouped into class intervals. *Clustering*

Suppose out "object" is now a cluster at distance y. Let

n_c be the number of clusters and let $\overline{c}$ = average cluster

size. Then, since we usually have $E[n_c|\overline{c}] = E[n_c]$,

$$\begin{aligned} E[n] &= E[n_c\overline{c}] \\ &= E_{\overline{c}}E[n_c\overline{c}|\overline{c}] = E_{\overline{c}}[\overline{c}E[n_c|\overline{c}]] \\ &= E_{\overline{c}}[\overline{c}E[n_c]] \\ &= E[\overline{c}]E[n_c] \\ &= E[c]E[n_c]\ . \end{aligned}$$

We then apply the above theory to the clusters.

2 Some models

There are two types of models for the detection function

$g(y)$, namely parametric and nonparametric. What would be a reasonable shape of $g(y)$? (Discuss - see handout).

2.1 Parametric models

We assume for the moment, that the detections are independent. (When does this break down?). Consider the half-normal model,

$$g(y) = e^{-\frac{1}{2\sigma^2}y^2}, \qquad 0 \le y < \infty \; .$$

Assuming no truncation ($w = \infty$),

$$\begin{aligned}\lambda &= \int_0^\infty g(y)dy = \int_0^\infty e^{-\frac{1}{2\sigma^2}y^2}dy \\ &= \sqrt{2\pi}\sigma \cdot \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}y^2}dy \\ &= \sigma\sqrt{\frac{\pi}{2}} \; .\end{aligned}$$

[Usually w is large enough to be able to assume that

$$\int_0^w g(y)dy \approx \int_0^\infty g(y)dy \; .$$

This is certainly the case in the half-normal model with $w \ge 4$.] We need to estimate σ. Now we have the joint distribution of all the random variables, namely

$$h(y_1, y_2, \ldots, y_n, n) = h_1(y_1, y_2, \ldots, y_n|n)h_2(n) \; .$$

In general, as we shall see below, $h_2(n)$ provides no information about σ. We therefore consider

$$\begin{aligned}L(\sigma) &= h_1(y_1, y_2, \cdots, y_n|n) \\ &= \prod_{i=1}^n f(y_i) \\ &= \prod_{i=1}^n \frac{g(y_i)}{\lambda} \\ &= \lambda^{-n} e^{-\frac{1}{2\sigma^2}\sum_i y_i^2} \; , \\ \log L(\sigma) &= -\frac{1}{2\sigma^2}\sum_i y_i^2 - n\log\sigma - n\log\sqrt{\frac{\pi}{2}} \; ,\end{aligned}$$

and

$$\frac{d \log L(\sigma)}{d\sigma} = \frac{\sum_i y_i^2}{\sigma^3} - \frac{n}{\sigma} = 0$$

implies that

$$\sigma = \left(\frac{1}{n} \sum_i y_i^2 \right)^{1/2} .$$

Thus

$$\hat{f}(0) = \frac{1}{\hat{\lambda}} = \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\hat{\sigma}} = \sqrt{\frac{2}{\pi \hat{\sigma}^2}} .$$

We now show that $h_2(n)$ provides no further information on σ. Since we are assuming that the animals are independent of one another, $n \sim \text{Binomial}(N, P)$, where from (2) $P = 2L\lambda/A$. Then

$$\begin{aligned} h(y_1, y_2, \cdots, y_n, n) &= h_1(y_1, y_2, \cdots, y_n | n) h_2(n) \\ &= \prod_{i=1}^{n} f(y_i) \binom{N}{n} P^n (1-P)^{N-n} , \end{aligned}$$

$$\frac{\partial \log h}{\partial \lambda} = \sum_{i=1}^{n} \frac{\partial \log f(y_i)}{\partial \lambda} + \left(\frac{n}{P} - \frac{N-n}{1-P} \right) \frac{\partial P}{\partial \lambda} = 0 ,$$

and

$$\Delta_N \log h = \log N - \log(N-n) + \log(1-P) = 0 .$$

This last equation implies that

$$\frac{N-n}{N} = 1 - P, \quad \text{or} \quad P = \frac{n}{N} ,$$

namely

$$\left(\frac{n}{P} - \frac{N-n}{1-P} \right) = 0 .$$

Hence

$$\frac{\partial \log h}{\partial \lambda} = \sum_{i=1}^{n} \frac{\partial \log f(y_i)}{\partial \lambda} ,$$

so that $h_2(n)$ provides no information about λ. It uses the information about λ from $h(y_1, y_2, \cdots, y_n | n)$ to estimate N. The half-normal model is not a very flexible model as it contains only one parameter σ. However it is useful as a possible starting point.

2.2 Robust parametric (semiparametric) models.

These models take the form

$$g(y) = a(y)(1 + b(y_s)),$$

where y_s is y scaled. Here $a(y)$ is called the "key" function. Useful key functions are (i) uniform ($g(y) = 1$) (ii) half-normal (above) and (iii) hazard-rate model ($g(y) = 1 - exp(-(y/\sigma)^{-b})$). The function $b(y)$ is the "series" function with, generally, 1 to 3 terms in the series. Useful series functions are:

(i) Cosine : $\sum_{j=1}^{m} a_j \cos(j\pi y_s)$.
(ii) Hermite polynomial : $\sum_{j=1}^{m} a_j H_j(y_s)$.
(iii) Polynomial : $\sum_{j=1}^{m} a_j y_s^j$,

where $y_s = \frac{y}{w}$ or $\frac{y}{\sigma}$; m has to be determined empirically. Non-parametric models

One such method is the kernel method for estimating a probability density function

3 Variance estimation and confidence intervals

Lemma 2 Let $X_1, X_2, \cdots, X_n$ be independent random variables with a common mean θ and variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$. Then $E[\frac{S^2}{n}] = E[\frac{\sum (X_i - \overline{X})^2}{n(n-1)}] = var[\overline{X}]$

Proof : $var[\overline{X}] = var[\frac{1}{n}\sum_i X_i] = \frac{1}{n^2}\sum_i \sigma_i^2$ and

$$\begin{aligned}
E[\sum (X_i - \overline{X})^2] &= E[\sum (X_i - \theta - (\overline{X} - \theta))^2] \quad (E(\overline{X}) = \theta) \\
&= E[\sum (X_i - \theta)^2 - 2(\overline{X} - \theta)\sum (X_i - \theta) + n(\overline{X} - \theta)^2] \\
&= E[\sum (X_i - \theta)^2 - n(\overline{X} - \theta)^2] \\
&= \sum_i \sigma_i^2 - nVar[\overline{X}] \\
&= n^2 Var[\overline{X}] - nVar[\overline{X}] \\
&= n(n-1)Var[\overline{X}].
\end{aligned}$$

Lemma 3 Suppose in Lemma 2 the variance

σ_i^2 are known. Let $\overline{X}_w = \sum_{i=1}^{n} w_i X_i$ be an unbiased estimate of θ. Then $var[\overline{X}_w]$ is minimized when $w_i \propto 1/\sigma_i^2$.

Proof : Now

$$\theta = E[\overline{X}_w] = \sum_i w_i E[X_i] = \theta \sum_i w_i$$

so that $\sum_i w_i = 1$. We need to minimize

$$var[\overline{X}_w] = \sum_i w_i^2 var[X_i] = \sum_i w_i^2 \sigma_i^2$$

subject to $\sum w_i = 1$.

The neatest way is to use a Lagrange multiplier, or else substituting $w_n = 1 - \sum_{i=1}^{n-1} w_i$ we minimize

$$v = \sum_{i=1}^{n-1} w_i^2 \sigma_i^2 + (1 - \sum_{i=1}^{n-1} w_i)^2 \sigma_n^2 \ .$$

Now

$$\frac{\partial v}{\partial w_i} = 0 \quad \Rightarrow \quad 2w_i\sigma_i^2 - 2(1 - \sum_{i=1}^{n-1} w_i)^2 \sigma_n^2 = 0$$

so that

$$2w_i\sigma_i^2 = 2w_n\sigma_n^2 \quad \text{for } i = 1, 2, \cdots, n-1.$$

Since $w_i\sigma_i^2 = a$, say, $w_i = a/\sigma_i^2$.

We now apply the above theory to our line transect theory. From (5), we have the estimate

$$\widehat{var}[\hat{D}] \simeq \hat{D}^2 \left\{ \frac{\hat{v}ar[\hat{f}(0)]}{\{\hat{f}(0)\}^2\}} + \frac{\hat{v}ar[n]}{n^2} \right\} = v_{\hat{D}} \ , \text{say.}$$

We can use a model-based estimate of $var(n)$ (e.g assume n is Poisson) or better, use replication. Suppose we have k line transects each of length l, and let n_j be the number observed from the jth transect.

Method 1: We combine all the data so that we effectively have a single transect of length $L = kl$. We proceed as before except that we now have an empirical estimate of $var[n]$, where $n = \sum_j n_j$. Thus

$$Var[n] = Var[\sum_j n_j] = k^2 var[\overline{n}], \quad \overline{n} = \frac{\sum n_j}{k} \ .$$

By Lemma 2, an unbiased estimate of $var[n]$ is therefore

$$v_n = \frac{k^2 \sum_{j=1}^{k} (n_j - \overline{n})^2}{k(k-1)} = \frac{k}{k-1} \sum_{j=1}^{k} (n_j - \overline{n})^2 \ .$$

The rest of $v_{\hat{D}}$ is obtained from the Distance computer package.

Method 2: Let

$$\hat{D}_j = \frac{n_j \hat{f}_j(0)}{2l}$$

be the estimate of D obtained from just the jth transect. We can now define

$$\overline{D} = \frac{1}{k}\sum_{j=1}^{k} \hat{D}_j$$

and estimate $var[\overline{D}]$ by the unbiased estimate

$$v_{\overline{D}} = \sum_j \frac{(\hat{D}_j - \overline{D})^2}{k(k-1)} \ .$$

Which is better, method 1 or method 2 ? Method 1 makes more use of the model structure : $v_{\hat{D}}$ will generally be smaller than $v_{\overline{D}}$. Method 2 may be more robust.

What happens if the transects are of different lengths, e.g. the jth has length l_j?

Method 1: Let $L = \sum_{j=1}^{k} l_j$. Now from equation (3) (Section 2)

$$E[n_j] = 2Dl_j\lambda \quad \text{and} \quad E\left[\frac{n_j}{l_j}\right] = 2D\lambda \ .$$

Let $z_j = n_j/l_j$, then $E[z_j] = \theta$, i.e. a common mean. Assuming the Poisson approximation, namely $var[n_j] = E[n_j]$, we have

$$var[z_j] = \frac{1}{l_j^2} Var[n_j] = \frac{2D\lambda}{l_j} \ .$$

The weighted average $\sum w_j z_j$ will have minimum variance when $w_j \propto l_j$; in fact when $w_j = l_j/L$. This weighted average is

$$\overline{z}_w = \sum_j \frac{l_j}{L}\frac{n_j}{l_j} = \frac{n}{L},$$

which is just what we want! Since $var[\frac{n}{L}] = \frac{1}{L^2} var[n]$, we can use Lemma 3 and obtain the following unbiased variance estimate :

$$\begin{aligned} \widehat{var}[n] &= L^2 \sum_j \frac{w_j(z_j - \overline{z}_w)^2}{(k-1)} \\ &= \frac{L\sum_j l_j \left(\frac{n_j}{l_j} - \frac{n}{L}\right)^2}{k-1} \ . \end{aligned}$$

Method 2:

$$var[\hat{D}_j] = D^2\left[\frac{var[\hat{f}_j(0)]}{\{f(0)\}^2} + \frac{var[n_j]}{E[n_j]^2}\right] \ .$$

We make the rough assumption that

$$var[\hat{f}_j(0)] \approx \frac{\sigma_j^2}{E[n_j]}$$

for some σ_j^2, and assume the Poisson model. Then

$$var[\hat{D}_j] \approx D^2 \left[\frac{\sigma^2}{f^2(0)E[n_j]} + \frac{1}{E[n_j]} \right]$$
$$= \frac{D^2}{2Dl_j\lambda} \left[\frac{\sigma_j^2}{f^2(0)} + 1 \right] \propto \frac{1}{l_j} .$$

Thus we have

$$\overline{D}_w = \sum_j l_j \hat{D}_j / L \quad (\text{with } E[\hat{D}_j] = D)$$

and

$$v_{\overline{D}_w} = \frac{\sum_j l_j (\hat{D}_j - \overline{D}_w)^2}{L(k-1)} .$$

The above theory is only approximate. Alternative methods such as bootstrapping are available, especially when k is small. There is also a theory based on the radial distance r.

4 Adaptive Line Transect

Adaptive sampling (see, Seber and Salehi(2002)) offers a means of increasing sample size, and hence increasing precision, by concentrating survey effort where most observations occur. Standard adaptive sampling methods can readily be extended to distance sampling surveys. For example, for point transect sampling we can define a grid of points, randomly superimposed on the study region, and randomly or systematically sample from the grid to form the primary sample. When a detection is made at a primary sample point, points from the grid that surround the primary sample point are sampled. If detections are made at these extra points, then further sampling is triggered. A major practical problem of adaptive sampling is that the required survey effort is not known in advance. This is particularly problematic for shipboard surveys, in which the ship is available for a predetermined number of days. Pollard and Buckland (1997) developed an adaptive sampling in shipboard line transect survey. The survey afford is increased when the number of observation exceeds some limit. The increased effort is achieved by zigzagging for a period, after which the ship returns to the nominal (straight line) cruse track. Unlike standard adaptive sampling, the method is not design-unbiased, but simulations indicate that the bias is small. An experimental trial on a survey of harbor porpoise in the Gulf of Maine yielded substantially more detections and better precision than did conventional line transect sampling.

References

Buckland, S.T., Anderson, D.R., Burnham, K.P. and Laake, J.L. (1993). *Distance Sampling: Estimating Abundance of Biological Populations*, Chapman & Hall, London, reprinted (1999) by Research Unit for Wildlife Population Assessment, St Andrews.

Eberhart, L.L.(1968). a preliminary appraisal of line transect. it J. Wildl. Manag. **32**, 82-88.

Gates,C.E., Marshall, W.H. and Olson, D.P. (1968). Line transect method of estimating grouse population densities. *Biometrics* **24**, 135-145.

Pollard, J.H. and Buckland, S.T. (1997). A Strategy for Adaptive Sampling in Shipboard Line Transect Surveys, Report of *the International Whaling Commission*, **47**, 356368.

Seber amd Salehi (2002). Adaptive Sampling, *Encyclopedia of Biostatistics*. Editors Peter Armitage and Theodore Colton. John Wiley & Sons, Ltd, Chichester, in press.

Admissible Estimation in an One Parameter Nonregular Family of Absolutely Continuous Distributions

Sanjari Farsipour, N.

P11080

Department of Statistics, Shiraz University, Iran.

Abstract. Consider the problem of estimating under entropy loss an arbitrarily positive, strictly increasing or decreasing parametric function based on a sample of size n in an one parameter noregular family of absolutly continuous distributions with both endpoints of the support depending on a single parameter. We first provide sufficient conditions for the admissibility of generalized Bayes estimtor with respect to some specific priors and then treat several examples which illustrate the admissibility of best invariant estimators is some location or scale parameter problems.

Keywords. One Parameter Nonregular Family, Generalized Bayes Estimator, Admissibility, Entropy Loss, Best Invariant Estimator.

1 Introduction

Karlin (1958) developed a method for proving the admissibility of estimators under squared error loss in the one parameter exponential family. Pulskamp and Ralescu (1991) gave sufficient conditions for the admissibility of nonlinear estimators in estimating an arbitrary parametric function using Zidek's (1970) formal Bayes approach as well as Karlin's method. In this paper we consider, using Karlins's method, the admissible estimation in an one parameter nonregular family of absolutely continuous distributions, when the loss is entropy loss function of the form

$$L(\theta,\delta)=\frac{\delta}{\theta}-ln\frac{\delta}{\theta}-1. \tag{1.1}$$

Let X have the density of the form

$$p(x;\theta)=\begin{cases} r(x)q(\theta)\ , a(\theta)<x<b(\theta); \theta\in(\underline{\theta},\bar{\theta}) \\ 0 \qquad\qquad \text{otherwise,} \end{cases} \tag{1.2}$$

with respect to lebesgue measure where $(\underline{\theta},\bar{\theta})$ is a nondegenerate interval in the real line which may be an infinite interval, $r(x)$ is a positive lebesgue measurable function of x,

$$q^{-1}(\theta)=\int_{a(\theta)}^{b(\theta)} r(x)dx<\infty \quad \text{for all} \quad \theta\in(\underline{\theta},\bar{\theta}),$$

and both $a(\theta)$ and $b(\theta)$ are functions of θ such that $a(\theta)<b(\theta)$ for all $\theta\in(\underline{\theta},\bar{\theta})$. For $a(\theta)=\underline{\theta}$ and $b(\theta)=\theta$, Karlin (1958) gave a single admissible estimator $\delta(X)=[(2\alpha+1)/(\alpha+1)]q^{-\alpha}(X)$ of $q^{-\alpha}(\theta),\alpha>0$, under squared error loss when $\lim_{\theta\to\bar{\theta}} q(\theta)=0$. For more general results in this problem see Sinha and Das Gupta (1984). Also, for $a(\theta)=\theta$ and $b(\theta)=\bar{\theta}$, Karlin (1958) provided a single admissible

estimator $\delta(X) = [(2\alpha+1)/(\alpha+1)]q^{-\alpha}(X)$ of $q^{-\alpha}(\theta), \alpha > 0$, under squared error loss when $\lim_{\theta \to \underline{\theta}} q(\theta) = 0$.

In this paper we treat the case when both $a(\theta)$ and $b(\theta)$ are strictly increasing functions of θ, and the loss is entropy loss (1.1). Suppose on the basis of random sample $X_1, X_2, \ldots, X_n$ of size $n(\geq 2)$ from the density (1.2) it is desired to estimate an arbitrarily positive, strictly increasing or decreasing function $h(\theta)$ under entropy loss (1.1). Let $X_{(1)}$ and $X_{(n)}$ be respectively the smallest and the largest members in a sample $X_1, X_2, \ldots, X_n$. Then the joint density of a sample $X_1, X_2, \ldots, X_n$ is given by

$$p(x_1, \ldots, x_n; \theta) = q^n(\theta)U(x_{(1)} - a(\theta))U(b(\theta) - x_{(n)})\Pi_{i=1}^n r(x_i),$$

where $U(y) = 1$ if $y \geq 0$ and $U(y) = 0$ otherwise. It follows from the factorization theorem that $X_{(1)}$ and $X_{(n)}$ are a pair of sufficient statistics of θ. Furthermore, it is well known (Kendall and Stuart (1979)) that there exists no single sufficient statistic, but $X_{(1)}$ and $X_{(n)}$ are jointly minimal suffieient for θ. Moreover, the strict convexity of the loss function guarantees that (from the viewpoint of risk) only nonrandomized estimators based on a (possibly minimal) sufficient statistic need be considered (see Berger (1985), p 40 -41). Consider the (possibly improper) prior of the form

$$\Pi_f(\theta) = \frac{|h'(\theta)|f(h(\theta))}{q^n(\theta)} \tag{1.3}$$

for almost all $\theta \in (\underline{\theta}, \bar{\theta})$ where f is a nonnegative function defined on the range of h. The prior under consideration is assumed to be absolutely continuous with respect to lebesgue measure with the density (1.3).

In Section 2 we provide, using Karlin's (1958) method, we prove the admissibility of the (nonrandomized) generalized Bayes estimator, of an arbitrarily positive, strictly increasing or decreasing function $h(\theta)$ with respect to the prior (1.3). Finally, Section 3 contains some examples for the results of Section 2.

2 Admissibility of the Generalized Bayes Estimator

Let $X_1, X_2, \ldots, X_n$ be a random sample from the density (1.2) where both $a(\theta)$ and $b(\theta)$ are strictly increasing functions of θ. It is understood that $\underline{\theta} \geq \eta$ where η is a unique value of θ such that $a(\eta) = b(\eta)$. Note that η may be $-\infty$. Also it is assumed that the range $(\underline{\theta}, \bar{\theta})$ of the parameter θ is wide enough so that $a(\bar{\theta}) > b(\underline{\theta})$.

First, consider the problem of estimating a positive, strictly increasing function $h(\theta)$ of θ. Then, the (nonrandomized) generalized Bayes estimator δ_f of $h(\theta)$ with respect to the prior (1.3) is given by

$$\delta_f(X_{(1)}, X_{(n)}) = \frac{\int_{\max\{h(b^{-1}(X_{(n)})), h(\underline{\theta})\}}^{\min\{h(a^{-1}(X_{(1)})), h(\bar{\theta})\}} f(u)du}{\int_{\max\{h(b^{-1}(X_{(n)})), h(\underline{\theta})\}}^{\min\{h(a^{-1}(X_{(1)})), h(\bar{\theta})\}} \frac{1}{u} f(u)du} \tag{2.1}$$

under certain integrability conditios imposed on f for δ_f to be well-defined. Next, in the problem of estimating any positive, strictly decreasing function $h(\theta)$ of θ, the(nonrandomized) generalized Bayes estimator δ_f of $h(\theta)$ with respect to the prior (1.3) is given by

$$\delta_f(X_{(1)}, X_{(n)}) = \frac{\int_{\max\{h(a^{-1}(X_{(1)})), h(\overline{\theta})\}}^{\{\min\{h(b^{-1}(X_{(n)})), h(\underline{\theta})\}} f(u)du}{\int_{\max\{h(a^{-1}(X_{(1)})), h(\bar{\theta})\}}^{\min\{h(b^{-1}(X_{(n)})), h(\underline{\theta})\}} \frac{1}{u} f(u)du} \tag{2.2}$$

under certain integrability conditions imposed on f for δ_f to be well-defined. The following theorem provides the admissibility of the generalized Bayes estimator (2.1) of any positive, strictly increasing function $h(\theta)$ with respect to the prior (1.3) under the entropy loss (1.1).

Theorem 2.1. Let $f \geq 0$ defined on $(0, \infty)$ be such that

$$\int_a^b f(u)du < \infty \quad \text{and} \quad \int_a^b \frac{1}{u} f(u)du < \infty \tag{2.3}$$

for every $0 < a < b < \infty$. Then, the generalized Bayes estimator (2.1) of $h(\theta)$ with respect to the prior (1.3) under entropy loss (1.1) is admissible.

Proof. If δ_f is not admissible, then there exists another estimator $\delta^{'}$ such that

$$\begin{aligned} &E_\theta \left[\frac{\delta^{'}(X_{(1)}, X_{(n)})}{h(\theta)} - ln \frac{\delta^{'}(X_{(1)}, X_{(n)})}{h(\theta)} - 1 \right] \\ &\geq E_\theta \left[\frac{\delta_f(X_{(1)}, X_{(n)})}{h(\theta)} - ln \frac{\delta_f(X_{(1)}, X_{(n)})}{h(\theta)} - 1 \right] \end{aligned} \tag{2.4}$$

for all $\theta \in (\underline{\theta}, \bar{\theta})$ with strict inequality for at least one θ. Note that the expectations in (2.4) operates through the joint density of $X_{(1)}$ and $X_{(n)}$ which is given by

$$\begin{aligned} &p_{X_{(1)}, X_{(n)}}(x_{(1)}, x_{(n)}; \theta) \\ &= n(n-1)q^n(\theta)[\textstyle\int_{x_{(1)}}^{x_{(n)}} r(x)dx]^{n-2} r(x_{(1)}) r(x_{(n)}) \end{aligned}$$

for $a(\theta) < x_{(1)} \leq x_{(n)} < b(\theta); \underline{\theta} < \theta < \bar{\theta}$. Now (2.4) implies

$$\begin{aligned} &\int\int_{a(\theta) < x_{(1)} \leq x_{(n)} < b(\theta)} \left(\frac{\delta^{'}(x_{(1)}, x_{(n)})}{\delta_f(x_{(1)}, x_{(n)})} - ln \frac{\delta^{'}(x_{(1)}, x_{(n)})}{\delta_f(x_{(1)}, x_{(n)})} - 1 \right) \\ &q^n(\theta) \left[\textstyle\int_{x_{(1)}}^{x_{(n)}} r(x)dx \right]^{n-2} r(x_{(1)}) r(x_{(n)}) dx_{(1)} dx_{(n)} \\ &\leq \int\int_{a(\theta) < x_{(1)} \leq x_{(n)} < b(\theta)} \left(\frac{\delta^{'}(x_{(1)}, x_{(n)})}{\delta_f(x_{(1)}, x_{(n)})} - \frac{\delta^{'}(x_{(1)}, x_{(n)})}{h(\theta)} + \frac{\delta_f(x_{(1)}, x_{(n)})}{h(\theta)} - 1 \right) \\ &q^n(\theta) \left[\textstyle\int_{x_{(1)}}^{x_{(n)}} r(x)dx \right]^{n-2} r(x_{(1)}) r(x_{(n)}) dx_{(1)} dx_{(n)}, \end{aligned} \tag{2.5}$$

for all $\underline{\theta} < \theta < \bar{\theta}$. Let θ_1 and θ_2 be such that $\underline{\theta} < \theta_1 < \theta_2 < \overline{\theta}$ and $b(\theta_1) < a(\theta_2)$ (Note that, without loss of generality, we can assume $b(\theta_1) < a(\theta_2)$ since if $b(\theta_1) \geq a(\theta_2)$, and $\theta_1 \to \underline{\theta}, \theta_2 \to \overline{\theta}$, then we have $b(\underline{\theta}) \geq a(\overline{\theta})$ which contradicts our assumption $a(\bar{\theta}) > b(\underline{\theta})$. Integrating both sides of (2.5) over (θ_1, θ_2) with respect to $\pi_f(\theta)$ in

(1.3) and then applying Fubini's theorem yield, after some calculations, the RHS of (2.5) as

$$\begin{array}{l}\int\int_{a(\theta_1)<x_{(1)}\leq x_{(n)}<b(\theta_1)} \\ \left[\int_{\theta_1}^{a^{-1}(x_{(1)})}\left(\frac{\delta^{'}}{\delta_f}-\frac{\delta^{'}}{h(\theta)}+\frac{\delta_f}{h(\theta)}-1\right)q^n(\theta)\Pi_f(\theta)d\theta\right] \\ \left[\int_{x_{(1)}}^{x_{(n)}} r(x)dx\right]^{n-2} r(x_{(1)})r(x_{(n)})dx_{(1)}dx_{(n)} \\ +\int\int_{a(\theta_2)<x_{(1)}\leq x_{(n)}<b(\theta_2)} \\ \left[\int_{b^{-1}_{(x_{(n)})}}^{\theta_2}\left(\frac{\delta^{'}}{\delta_f}-\frac{\delta^{'}}{h(\theta)}+\frac{\delta_f}{h(\theta)}-1\right)q^n(\theta)\Pi_f(\theta)d\theta\right] \\ \left[\int_{x_{(1)}}^{x_{(n)}} r(x)dx\right]^{n-2} r(x_{(1)})r(x_{(n)})dx_{(1)}dx_{(n)}\end{array} \tag{2.6}$$

where $\delta_f \equiv \delta_f(x_{(1)}, x_{(n)})$ and $\delta^{'} \equiv \delta^{'}(x_{(1)}, x_{(n)})$. Substituting for $\Pi_f(\theta)$, and simplifying, (2.6) becoms

$$\begin{array}{l}\int_{a(\theta_1)}^{b(\theta_1)}\int_{a(\theta_1)}^{x(n)}\left(\frac{\delta^{'}}{\delta_f}-1\right)\int_{\theta_1}^{a^{-1}(x_{(1)})} f(x)dx\left[\int_{x_{(1)}}^{x_{(n)}} r(x)dx\right]^{n-2} r(x_{(1)})r(x_{(n)})dx_{(1)}dx_{(n)} \\ +\int_{a(\theta_1)}^{b(\theta_1)}\int_{a(\theta_1)}^{x(n)}\left(\delta_f-\delta^{'}\right)\int_{\theta_1}^{a^{-1}(x_{(1)})}\frac{1}{x}f(x)dx\left[\int_{x_{(1)}}^{x_{(n)}} r(x)dx\right]^{n-2} r(x_{(1)})r(x_{(n)})dx_{(1)}dx_{(n)} \\ +\int_{a(\theta_2)}^{b(\theta_2)}\int_{a(\theta_2)}^{x_{(n)}}(\frac{\delta^{'}}{\delta_f}-1)\int_{b^{-1}(x_{(n)})}^{\theta_2} f(x)dx\left[\int_{x_{(1)}}^{x_{(n)}} r(x)dx\right]^{n-2} r(x_{(1)})r(x_{(n)})dx_{(1)}dx_{(n)} \\ +\int_{a(\theta_2)}^{b(\theta_2)}\int_{a(\theta_2)}^{x_{(n)})}(\delta_f-\delta^{'})\int_{b^{-1}(x_{(n)})}^{\theta_2}\frac{1}{x}f(x)dx\left[\int_{x_{(1)}}^{x_{(n)}} r(x)dx\right]^{n-2} r(x_{(1)})r(x_{(n)})dx_{(1)}dx_{(n)}\end{array} \tag{2.7}$$

Now, letting $\theta_1 \to \underline{\theta}$ and $\theta_2 \to \overline{\theta}$ and using the assumption (2.3), we can see that (2.7) tends to zero, which shows that δ_f is admissible. The following theorem gives the admissibility of the generalized Bayes estimator (2.2) of any positive, strictly decreasing function $h(\theta)$ with respect to the prior (1.3). The proof is omitted because it parallels that of Theorem 2.1.

Theorem 2.2. Let $f \geq 0$ defined on $(0, \infty)$ be such that

$$\int_a^b f(u)du < \infty \quad \text{and} \quad \int_a^b \frac{1}{u}f(u)du < \infty \tag{2.8}$$

for every $0 < a < b < \infty$. Then the generalized Bayes estimator (2.2) of $h(\theta)$ with respect to the prior (1.3) under entropy loss (1.1) is admissible.

3 Examples

In this Section we present several examples for Theorem 2.1 and Theorem 2.2. In the following examples we consider $f(u) = u^{-\alpha}; u > 0, \alpha > 1$ in Theorem 2.1 and $f(u) = u^{\alpha+1}; u > 0, \alpha < -1$ in Theorem 2.2, where they satisfy conditions (2.3) and (2.8) of Theorems 2.1 and 2.2 respectively.

Example 3.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from the density

$$p(x;\theta) = \begin{cases} 1/\theta \ s\theta < x < (s+1)\theta, s > 0(\text{known}), 0 < \theta < \infty \\ 0 \quad \text{otherwise.} \end{cases}$$

Here, $a(\theta) = s\theta$, $b(\theta) = (s+1)\theta$, $\eta = \underline{\theta} = 0, \bar{\theta} = \infty, r(x) \equiv 1$, and $q(\theta) = \frac{1}{\theta}$. Note that $a(\bar{\theta}) > b(\underline{\theta})$. Now, we want to estimate $h(\theta) = \theta^k, k > 0$ under entropy loss (1.1). With $f(u) = u^{-\alpha}, u > 0, \alpha > 1$, in Theorem 2.1 the prior in (1.3) is $\Pi_f(\theta) \equiv \Pi_\alpha(\theta) = k\theta^{n-k(\alpha-1)-1}; \theta > 0, \alpha > 1$, and for this prior, the generalized Bayes estimator $\delta_f \equiv \delta_\alpha$ in (2.1) is

$$\delta_\alpha(X_{(1)}, X_{(n)}) = \frac{\alpha}{\alpha - 1}\left[\frac{(\frac{X_{(n)}}{s+1})^{-k(\alpha-1)} - (\frac{X_{(1)}}{s})^{-k(\alpha-1)}}{(\frac{X_{(n)}}{s+1})^{-\alpha k} - (\frac{X_{(1)}}{s})^{-\alpha k}}\right],$$

which is admissible by Theorem 2.1.

Remark 3.1. Note that all moments are multiples of θ^k for appropriate k's, e.g., mean $\left(= \frac{s+1}{2}\theta\right)$ and the variance $(= (1/12)\theta^2)$ are multiples of θ and θ^2, respectively. Also, the quantiles $\xi_p = \theta(s+p), 0 < p < 1$, are multiples of θ. It can be shown that $\delta_{n+1}(X_{(1)}, X_{(n)})$ for k=1, s=1 in Example 3.1 is the best invariant estimator (Pitman's estimator) of the scale parameter θ under the entropy loss function (1.1), and is admissible by Example 3.1, under loss (1.1).

Example 3.2. Let $X_1, X_2, \ldots, X_n$ be as in Example 3.1. Suppose it is desired to estimate $h(\theta) = \theta^k, k < 0$, under entropy loss (1.1). With $f(u) = u^{\alpha+1}; u > 0, \alpha < -1$, in Theorem 2.2, the prior (1.3) becomes $\Pi_f(\theta) \equiv \Pi_\alpha(\theta) = |k|\theta^{n+k(\alpha+2)-1}; \theta > 0$, and the generalized Bayes estimator $\delta_f \equiv \delta_\alpha$ in (2.2) with respect to this prior is

$$\delta_\alpha(X_{(1)}, X_{(n)}) = \frac{\alpha - 1}{\alpha}\left[\frac{(\frac{X_{(n)}}{s+1})^{k\alpha} - (\frac{X_{(1)}}{s})^{k\alpha}}{(\frac{X_{(n)}}{s+1})^{k(\alpha-1)} - (\frac{X_{(1)}}{s})^{k(\alpha-1)}}\right]$$

which is admissible under entropy loss (1.1) by Theorem 2.2.

Example 3.3. Let $X_1, X_2, \ldots, X_n$ be a random sample from the density

$$p(x;\theta) = \begin{cases} 1 \ , \theta < x < \theta + 1, -\infty < \theta < \infty \\ 0 \text{ otherwise} \end{cases}$$

In this case, $a(\theta) = \theta, b(\theta) = \theta + 1, q(\theta) \equiv 1, r(x) \equiv 1, \underline{\theta} = -\infty, \bar{\theta} = \infty$, and η is taken to be $-\infty$. Note that $a(\bar{\theta}) > b(\underline{\theta})$. Now, we want to estimate $h(\theta) = e^{t\theta}$ under entropy loss (1.1), where $t \neq 0$ is a real number. Note that the moment generating function $\{(e^t - 1)/t\}e^{t\theta}$ of X_1 is a multiple of $h(\theta) = e^{t\theta}$. With $t > 0$ and $f(u) = u^{-2}; u > 0$, in Theorem 2.1, the prior (1.3) becomes $\Pi_f(\theta) = te^{-t\theta}, -\infty < \theta < \infty$, and the generalized Bayes estimator $\delta_f \equiv \delta$ in (2.1) with respect to this prior is given by;

$$\delta(X_{(1)}, X_{(n)}) = \frac{1}{2}\left[\frac{e^{-tX_{(1)}} - e^{-t(X_{(n)}-1)}}{e^{-2tX_{(1)}} - e^{-2t(X_{(n)}-1)}}\right]$$

which is admissible for estimating $e^{t\theta}, t > 0$ under entropy loss (1.1) by theorem (2.1).

Remark 3.2. As a special case of Example 3.3 consider the case when $t = 1$. Then we have admissibility of

$$\delta(X_{(1)}, X_{(n)}) = \frac{1}{2}\left[\frac{e^{-X_{(1)}} - e^{-(X_{(n)}-1)}}{e^{-2X_{(1)}} - e^{-2(X_{(n)}-1)}}\right]$$

for estimating $h(\theta) = e^{\theta}$ under entropy loss (1.1). In fact, this estimator is the best invariant estimator of e^{θ} under Linex loss of the form

$$L(\delta, \theta) = e^{(\delta-\theta)} - (\delta - \theta) - 1$$

Example 3.4. Let $X_1, X_2, \ldots, X_n$ be as in Example 3.3. We want to estimate $h(\theta) = e^{t\theta}, t < 0$. With $f(u) = u^{-2}; u > 0$ in Theorem 2.2 the prior $\Pi_f = |t|e^{-t\theta}; -\infty < \theta < \infty$, and the generalized Bayes estimator $\delta_f \equiv \delta$ in (2.2) is given by

$$\delta(X_{(1)}, X_{(n)}) = \frac{1}{2}\left[\frac{e^{-tX_{(1)}} - e^{-t(X_{(n)}-1)}}{e^{-2tX_{(1)}} - e^{-2t(X_{(n)}-1)}}\right]$$

and is admissible for estimating $e^{t\theta}; t < 0$ under entropy loss (1.1). Combining this result with that of Example 3.3 yields that the generalized Bayes estimator

$$\delta(X_{(1)}, X_{(n)}) = \frac{1}{2}\left[\frac{e^{-tX_{(1)}} - e^{-t(X_{(n)}-1)}}{e^{-2tX_{(1)}} - e^{-2t(X_{(n)}-1)}}\right]$$

is admissible for estimating $e^{t\theta}; t \in R(t \neq 0)$.

References

1. Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis, 2nd ed., Springer-Verlag, New York.
2. Karlin, S. (1958). Admissibility for estimation with quadratic loss. Ann. Math. Statist. 29, 409-436.
3. Kendall, M. G. and Stuart, A. (1979). The Advanced theory of Statistics, Vol. 2, 4th ed. Chales Griffin, London.
4. Pulskamp, R. J. and Ralesucu, S. (1981). A general class of nonlinear admissible estimators in the one-parameter exponential case, J. Statist. planning and Inference, 28, 383-390.
5. Sharma, D. (1975). A note on Karlin's admissible result for extreme value theory, J. Ind. Statist. Assoc, 13, 67-69.
6. Singh, R. (1971). Admissible estimators of θ^r in some extreme value densities, Canad. Math. Bull., 14(3), 411-414.
7. Sinha, B.K. and Das Gupta, A. (1984). Admissibility of generalized Bayes and Pitman estimates in the nonregular family, Commun. Statist.- Theor. Meth., 13(4), 1709-1721.
8. Zidek, J. V. (1970). Sufficient conditions for the admissibility under squared errors loss of formal Bayes estimators, Ann. Math. Statist., 41, 446-456.

Estimation of the Scale Parameters in Continuous Populations

Sanjari Farsipour, N.

P21080

Department of Statistics, Shiraz University, Iran.

Abstract. Estimation of Scale Parameter with restrictions to the Principle of invariance under several loss function is considered. The minimum risk scale equivariant estimator is obtained, and the admissibility and inadmissibility classes are studied. The application to the multivariate case is considered.

Keywords. Admissibility, Asymmetric Loss, Bayes Estimator, Best Scale - Equivariant Estimator, Scale Parameter.

1 Introduction

Estimation of scale parameters was first studied by Pitman (1939). For a random sample $X_1, \ldots, X_n$ from a density $\frac{1}{\tau} f\left(\frac{x}{\tau}\right)$, the estimate for τ

$$\delta(x) = \frac{\int_0^\infty \upsilon^n \prod \lim_{i=1}^n f(\upsilon x_i)\, d\upsilon}{\int_0^\infty \upsilon^{n+1} \prod \lim_{i=1}^n f(\upsilon x_i)\, d\upsilon} \tag{1.1}$$

was shown to have a negative bias, although it possesses optimal properties among the class of estimators with the multiplicative property. The estimator $\delta(x)$ in the form (1.1) is known as the pitman estimator of τ when the loss is

$$L(\tau, \delta) = \left(\frac{\delta}{\tau} - 1\right)^2 \tag{1.2}$$

Pitman's work has been extended by Girshick and Savage (1951) and others in the direction of minimax estimation.

The admissibility of the best invariant estimator (1.1) for the variance of a normal distribution with known mean μ under the loss (1.2), was proved by Hodges and Lehmann (1951) and Girshick and Savage (1951). Also the inadmissibility of the best invariant estimator for the Normal variance when the mean μ is unknown was proved by Stein (1964) under the loss (1.2). See also Matta and Casella (1990).

Form decision theoretic approach when symmetrics are present in a problem, it is natural to require a corresponding symmetry to hold for the estimators. This strongly suggests that the statistician should use an estimation procedure which also has the property of being invariant.

Dealing with Scale parameters, with group

$$\mathcal{G} = \{g_c; g_c(x_1, \ldots, x_n) = (cx_1, \ldots, cx_n); c > 0\}$$

and any loss $L(\tau,\delta)$ of the form $\rho(\delta,\tau)$, the class of all scale - invariant estimators of τ is of the form $\delta(x)=\delta_0(x)/w(z)$, where δ_0 is any scale - invariant estimator, $X=(x_1,\ldots,x_n)$ and $z=(z_1,\ldots,z_n)$ with $z_i=\frac{X_i}{X_n}, i=1,....,n-1$ $z_n=\frac{Xn}{|X_n|}$. The best scale-invariant estimator $\delta^*$of τ is given by $\delta^*(X)=\frac{\delta_0(X)}{w^*(Z)}$ where $w^*(z)$ is a number which minimizes $E_{\tau=1}\{\rho(\delta_0(X)/w(z))|Z=z\}$. In the presence of a location parameter as a nuisance parameter, the best scale invariant estimator of τ, is of the form $\delta^*(X)=\delta_0(Y)/w^*(z)$, where $\delta_0(Y)$ is any finite risk scale-invariant estimator of τ, based on $Y=(Y_1,\ldots,Y_{n-1})$, where $Y_i=X_i-X_n \quad i=1,\ldots,n-1$ $z=(z_1,\ldots,z_{n-1})$ with $z_i=\frac{Y_i}{Y_{n-1}}, i=1,....,n-2$ $z_{n-1}=\frac{Y_{n-1}}{|Y_{n-1}|}$ and $w^*(z)$ is any number minimizing $E_{\tau=1}\{\rho(\delta_0(Y)/w(z))|Z=z\}$, see Lehmann (1983). A loss function $L(\delta,\tau)$ represents the amount by which a statistician is penalized when τ is the true state of nature and δ is the statistician's action. In the literature, $L(\delta,\tau)$ is usually taken to be convex in δ and even in $\delta-\tau$, such as (1.2). This loss function has been criticized by some researchers (e.g Rukhin and Ananda (1992), Dey, Ghosh and Sirinivasan (1987), Akaike (1977, 1978)). They motivated the entropy loss as an asymmetric loss function for estimating on unknown scale parameter τ which is of the form

$$L(\delta,\tau)=\frac{\delta}{\tau}-\ln\frac{\delta}{\tau}-1 \tag{1.3}$$

The loss (1.3) was first introduced in James and Stein (1961) for estimation of the multinormal variance - covariance matrix. Later the same loss was considered in Brown (1968), Haff (1982), for estimating either the multinormal variance - covariance matrix or its inverse. This loss is also known as Stein loss.

In practice the real loss function is often not symmetric, overestimation of a parameter can lead to more or less sever consequences than underestimation. Examples of such cases are: In food-processing industries it is undesirable to overfill Containers, since there is no cost recovery for the overfile. If the containers are underfilled, however, it is possible to incur a much more severe penalty arising from misrepresentation of the product's actual weight or volume. (see Harris (1992)). Other examples may be found in Kuo and Dey (1990), Schäbe (1992), Zellner (1986). Varian (1975) employed the asymmetric loss function in real estate assessment, which is given by

$$L(\Delta)=b\left\{e^{a\Delta}-a\Delta-1\right\}, \tag{1.4}$$

where $a\neq 0$ determine the shape of the loss function and $b>0$ serves to scale the loss function. It is suitable for scale parameter estimation if $\Delta=\left(\frac{\delta}{\tau}-1\right)$. These losses have infinite maximum value, and in discribing for example the loss associated with a product, we can use a loss function of the form

$$L(\delta,\tau)=b\left\{1-e^{a\left(2-\frac{\delta}{\tau}-\frac{\tau}{\delta}\right)}\right\}, \tag{1.5}$$

where $a>0$ is a shape parameter and $b>0$ is the maximum loss parameter. This is obviously a bounded loss function, which is appropriate for scale parameter estimation. Another proposed loss function, which a bounded asymmetric loss for the scale parameter estimation is of the form

$$L(\delta,\tau)=k\left\{1-e^{b\left\{1+a\left(\frac{\delta}{\tau}-1\right)-e^{a\left(\frac{\delta}{\tau}-1\right)}\right\}}\right\}, \tag{1.6}$$

where $a\neq 0, b, k>0$. The loss (1.6) is scale invariant and bounded. $a\neq 0$ determine the shape of the loss function, $b>0$ serves to scale the loss and $k>0$ is the maximum loss parameter.

2 Best scale Invariant Estimator

Let $X_1,\ldots,X_n$be a random sample of size n from $\frac{1}{\tau}f(\frac{x}{\tau})$, where f is known and τ is unknown scale parameter. The joint density is denoted by

$$f_\tau(x)=\prod\lim_{i=1}^{n}\frac{1}{\tau}f(\frac{x_i}{\tau}) \tag{2.1}$$

In many cases the model (2.1) reduces to $C(x,n)\,\eta^{-\nu}e^{-T(x)/\eta}$, where $c(x,n)$ is a function of x and $n, \eta=\tau^r$, for some r,υ is a function of n and $T(X)$ is a complete sufficient statistic for η with $\Gamma(\nu,\eta)$-distribution. Examples of such models are: $\Gamma(\alpha,\beta)$ with α known and $\eta=\beta$; $E(0,\beta)$ with $\eta=\beta$; $N\left(0,\sigma^2\right)$ with $\eta=\sigma^2$; inverse Gaussian with zero drift and $\eta=\frac{1}{\lambda}$. Now, if $Z=(z_1,\ldots,z_n)$ with $z_i=\frac{X_i}{X_n}, i=1,\ldots,n-1, z_n=\frac{X_n}{|X_n|}$ and the loss is (1.3), and there exists a scale-equivariant estimator δ_0 of τ^r with finite risk, then a MRE estimator of τ^r is given by $\delta^*(X)=\frac{\delta_0(X)}{w^*(Z)}$, where $\omega^*(z)=\left\{E_{\tau=1}\left[\frac{1}{\delta_0(X)}\middle| Z=z\right]\right\}^{-1}$, and it can be shown that the MRE becomes $\delta^*(x)=\frac{\int_0^\infty \upsilon^{n-r-1}f(\nu x)d\nu}{\int_0^\infty \upsilon^{n-1}f(\nu x)d\nu}$, which is the generalized Bayes estimator of τ^r with respect to noninformative prior $\pi(\tau)=\frac{1}{\tau}$ under loss (1.3), which we refer to it as Pitman type estimator of τ^r. On the other hand suppose the loss is (1.4) with $\Delta=\frac{\delta}{\tau}-1$, and δ_0 is a scale-equivariant estimator of τ with finite risk, then the MRE estimator of τ is given by $\delta^*(x)=\delta_0(x)/w^*(z)$, where $w^*(z)$ is a solution of

$$E_{\tau=1}\left\{\delta_0(X)\,e^{\delta_0(x)/w^*(z)}\middle| Z=z\right\}=e^a E_{\tau=1}\left\{\delta_0(X)|\,Z=z\right\}. \tag{2.2}$$

In the case where the loss is (1.5) , with $a=b=1$, and under $\tau=1$, we can find (as mentioned) an equivariant estimator $\delta_0(X)$ or $\delta_0(Y)$ which has the gamma-distribution with known parameter υ,η and is independent of z. It follows that $\delta^*=\frac{\delta_0}{w^*}$ is the MRE estimator of τ where w^* is a number which maximizes

$$\begin{aligned} g(w)&=\int_0^\infty e^{2-\frac{x}{w}-\frac{w}{x}}\left\{\frac{\eta^\nu}{\Gamma(\nu)}x^{\nu-1}e^{-\eta x}\right\}dx\\ &=\frac{2e^2\eta^\nu w^\nu}{\Gamma(\upsilon)(1+\eta w)^{\upsilon/2}}k_\upsilon\left(2\sqrt{1+w\eta}\right)\\ &=C(\eta,\nu)\,\frac{w^\upsilon}{(1+\eta w)^{\upsilon/2}}k_\upsilon\left(2\sqrt{1+w\eta}\right), \end{aligned} \tag{2.3}$$

where C is a function of η and ν, and $k_f(.)$ is the modified Bessel function of order f (Gradshteyn and Ryzhik (1980)). So by using the relation $k_{v-1}(z) - k_{v+1}(z) = -\frac{2v}{z}k_\nu(z)$, the minimizing w i.e w^* must satisfy the follwoing equation

$$\frac{{w^*}^2}{1+w^*\eta}k_{v+1}\left(2\sqrt{1+w^*\eta}\right) = k_{v-1}\left(2\sqrt{1+w^*\eta}\right) \tag{2.4}$$

under the loss function (1.6) with $k = b = 1$, when $\tau = 1$ and we can find an equivariant estimator $\delta_0(X)$ or $\delta_0(Y)$ which has the gamma distribution with known parameters ν, η and is independent of Z, It follwos that $\delta^* = \frac{\delta_0}{w^*}$ is the MRE estimator of τ where w^* is a number which satisfying the following equation

$$\int_0^\infty x^{v-1}e^{\left(\frac{2a}{w^*}-\eta\right)x - e^{\frac{ax}{w^*}-a}}dx = e^a\int_0^\infty x^v e^{\left(\frac{2a}{w^*}-\eta\right)x - e^{\frac{ax}{w^*}-a}}dx \tag{2.5}$$

3 Bayes Estimation of Scale Parameter

The conjugate family of prior distributions for $\lambda = \eta^{-1}$ is the family of gamma distributions $\Gamma(\alpha,\beta)$, with density $\pi(\lambda|\alpha,\beta) = \beta^\alpha\lambda^{\alpha-1}e^{-\beta\lambda}/\Gamma(\alpha); \lambda > 0$, where $\alpha > 0, \beta > 0$. Note that the usual noninformative prior for λ is $\pi(\lambda) \propto \lambda^{-1}; \lambda > 0$, and corresponds to the limiting case $\alpha, \beta \longrightarrow 0$. Then the posterior distribution of λ is $\Gamma(\nu+\alpha, 1/(\beta+T(x)))$, and the Bayes estimator of η under (1.3) is $\delta_{Bayes}(X) = E(\eta|X) = E\left(\frac{1}{\lambda}|X\right) = \frac{T(X)+\beta}{\alpha+\nu-1}$, which can be written as

$$\delta_{Bayes}(X) = \frac{1}{v+\alpha-1}T(X) + \frac{\beta}{v+\alpha-1} = CT(X) + d \tag{3.1}$$

under the loss function (1.4) the unique Bayes estimator is of the form

$$\delta_{Bayes}(X) = c(\alpha)T(X) + d(\alpha,\beta) \tag{3.2}$$

where

$$c(\alpha) = \frac{1}{\alpha}\left(1 - e^{-\frac{a}{n+\alpha+1}}\right), d(\alpha,\beta) = c(\alpha)\beta,$$

provided $\beta + T(X) + a\delta_{Bayes}(X) > 0$ with considering (1.5) as a loss function, and letting $\delta_0(X)$ has $\Gamma\left(v, \frac{\eta}{\tau}\right)$-distribution, where $\upsilon > 0, \eta > 0$, with conjugate family of prior distribution for $\beta = \frac{1}{\tau}$ as a $\Gamma(\alpha,\xi)$, the posterior becomes $\Gamma(v+\alpha, \xi+\eta\delta_0(x))$, hence the Bayes estimator $\delta_{Bayes} = \delta_B$ must satisfies

$$k_{v+\alpha-1}\left(2\sqrt{\frac{1}{\delta_B}(\delta_B+\xi+\eta\delta_0(x))}\right) = \frac{\delta_B}{\delta_B+\xi+\eta\delta_0(x)}k_{v+\alpha+1}\left(2\sqrt{\frac{1}{\delta_B}(\delta_B+\xi+\eta\delta_0(x))}\right), \tag{3.3}$$

and the Bayes estimator for (1.6) must satisfies

$$\int_0^\infty \beta^{v+\alpha}e^{(2a\delta_B-\xi-\eta\delta_0(x))\beta - e^{a(\beta\delta_B-1)}}d\beta = e^a\int_0^\infty \beta^{v+\alpha}e^{(2a\delta_B-\xi-\eta\delta_0(x))\beta - e^{a(\beta\delta_B-1)}}d\beta \tag{3.4}$$

4 Admissibility and Inadmissibility Results

The estimator (3.1) is admissible, provided $0 \leq c < c^*, d > 0$ and $v > 1$. Where $c^* = 1/(v-1)$, and $c = c^*, d > 0$ and $v > 1$, and is inadmissible whenever one of the following conditions hold

(i) $c < 0$ or $d < 0$

(ii) $c = 0$ and $d = 0$

(iii) $0 < c \neq c^*$or $d = 0$

(iv) $c > c^*$ and $d = 0$.

The estimator (3.2) is admissible provided $0 \leq c < c^*$, $d > 0$, where $c^* = \frac{1}{a}\left(1 - e^{-\frac{a}{n+1}}\right)$, and is inamissible whenever one of the following conditions hold

(i) $c < 0$ or $d < 0$

(ii) $c > c^*$ and $d \geq 0$,

(iii) $0 \leq c < c^*$, $d = 0$.

Now if the loss is (1.3) with $\tau = \lambda^r$, where X belonges to the density

$$f(x, s, \lambda) = \frac{\lambda^s}{\Gamma(s)} x^{s-1} e^{-\lambda x}; x > 0, s > 0, \lambda \in \Lambda \tag{4.1}$$

and $\Lambda = (0, \lambda_0)$ or $\Lambda = (\lambda_0, \infty), \lambda_0$ is a given constant. $\lambda_0 \epsilon \Re_+ U \{\infty\}$. Admissible estimators of λ^r where obtained by Ghosh and Singh (1972). Using Karlin's method (cf. Karlin 1958) Ghosh and Singh (1970), proved admissibility of the estimator $(s-2)X^{-1}$ of the parameter λ. This result was generalized by Singh (1972) who showed that $\frac{\Gamma(s-r)}{\Gamma(s-2r)} X^{-1}$ is admissible estimator of λ^r under squared error loss, where r is an integer, $r < \frac{s}{2}$. Ghosh and Meeden (1977) and Ralescu and Ralescu(1981) hvae found admissible estimator of λ and λ^{-1} in the gamma distribution in the truncated parameter space under SEL. Let us denote $\gamma(.,.)$, $\Gamma(.,.)$ the incomplete gamma functions i.e.

$$\gamma(x, y) = \int_0^y t^{x-1} e^{-t} dt$$

and

$$\Gamma(x, y) = \int_y^\infty t^{x-1} e^{-t} dt \qquad\qquad x, y > 0$$

we now give an admissible estimator of λ^r under (1.3), where r is an integer $r < \frac{s}{2}$, and X is distributed according to (4.1).

Theorem 4.1: Suppose that the estimator

$$(i)\ \ \widehat{u}\,(X) = \frac{\gamma(s-r,\lambda_0(k+x))}{\gamma(s-2r,\lambda_0(k+x))}\,(x+k)^{-r}\ \ 0<\lambda<\lambda_0$$
$$(ii)\ \widehat{u}\,(X) = \frac{\Gamma(s-r,\lambda_0(k+x))}{\Gamma(s-2r,\lambda_0(k+x))}\,(x+k)^{-r}\ \ \lambda_0<\lambda<\infty$$

where $k \geq 0$ is an arbitary constant. Then $\widehat{u}$ is admissible for λ^r under entropy loss function (1.3).

Proof : Using Karlin's method we shall first prove case (i). Suppose that there exists an estimator $\widetilde{u}$ which is better that $\widehat{u}$. This implies that the inequality

$$\int_0^\infty \left(\frac{\widetilde{u}}{\lambda^r} - \ln\frac{\widetilde{u}}{\lambda^r} - 1\right) f\,(x,\lambda)\,dx$$
$$\leq \int_0^\infty \left(\frac{\widehat{u}}{\lambda^r} - \ln\frac{\widehat{u}}{\lambda^r} - 1\right) f\,(x,\lambda)\,dx$$

holds for all $\lambda \in \Lambda$ with strict inequality for some λ. After some calculations we get

$$\begin{array}{l}\int_0^\infty \left(\frac{\widetilde{u}}{u} - \ln\frac{\widetilde{u}}{u} - 1\right) f\,(x,\lambda)\,dx \\ \leq \int_0^\infty \left(\frac{\widetilde{u}}{u} - \frac{\widetilde{u}}{\lambda^r} + \frac{\widehat{u}}{\lambda^r} - 1\right) f\,(x,\lambda)\,dx\end{array} \tag{4.2}$$

Integrating both sides of (4.2) with respect to the improper prior

$$\xi\,(\lambda) = \lambda^{-2r-1}\exp\left(-k\lambda\right);\lambda \in (0,\lambda_0)\,,$$

we have

$$\begin{array}{l}\int_b^{\lambda_0}\int_0^\infty \left(\frac{\widetilde{u}}{u} - \ln\frac{\widetilde{u}}{u} - 1\right) f\,(x,\lambda)\,\xi\,(\lambda)\,dxd\lambda \\ \leq \int_b^{\lambda_0}\int_0^\infty \left(\frac{\widetilde{u}}{u} - \frac{\widehat{u}}{\lambda^r} + \frac{\widehat{u}}{\lambda^r} - 1\right) f\,(x,\lambda)\,\xi\,(\lambda)\,dxd\lambda\end{array} \tag{4.3}$$

Now interchanging the order of integration in the right hand side of (4.3) we have

$$\begin{array}{l}\int_0^\infty \frac{\widetilde{u}}{u\Gamma(s)} x^{s-1}\int_b^{\lambda_0}\lambda^{s-r-1}e^{-(x+k)\lambda}d\lambda dx \\ -\int_0^\infty \frac{u}{\Gamma(s)} x^{s-1}\int_b^{\lambda_0}\lambda^{s-2r-1}e^{-(x+k)\lambda}d\lambda dx \\ +\int_0^\infty \frac{u}{\Gamma(s)} x^{s-1}\int_b^{\lambda_0}\lambda^{s-2r-1}e^{-(x+k)\lambda}d\lambda dx \\ -\int_0^\infty \frac{1}{\Gamma(s)} x^{s-1}\int_b^{\lambda_0}\lambda^{s-r-1}e^{-(x+k)\lambda}d\lambda dx.\end{array} \tag{4.4}$$

Now substituting for $\widehat{u}$ in (4.4) we have

$$\begin{array}{l} \int_0^\infty \frac{\widehat{u\gamma(s-2r,\lambda_0(k+x))}}{\gamma(s-r,\lambda_0(k+x))\Gamma(s)}(x+k)^r x^{s-1}\int_b^{\lambda_0}\lambda^{s-r-1}e^{-(x+k)\lambda}d\lambda dx \\ -\int_0^\infty \frac{u}{\Gamma(s)}x^{s-1}\int_b^{\lambda_0}\lambda^{s-2r-1}e^{-(x+k)\lambda}d\lambda dx \\ +\int_0^\infty \frac{\gamma(s-r,\lambda_0(k+x))}{\gamma(s-2r,\lambda_0(k+x))\Gamma(s)}(x+k)^{-r}x^{s-1}\int_b^{\lambda_0}\lambda^{s-2r-1}e^{-(x+k)\lambda}d\lambda dx \\ -\int_0^\infty \frac{x^{s-1}}{\Gamma(s)}\int_b^{\lambda_0}\lambda^{s-r-1}e^{-(x+k)\lambda}d\lambda dx. \end{array} \tag{4.5}$$

Now, as $b \longrightarrow 0, \int_b^{\lambda_0}\lambda^{s-r-1}e^{-(x+k)\lambda}d\lambda \longrightarrow (k+x)^{r-s}\gamma(s-r,\lambda_0(k+x))$ and $\int_b^{\lambda_0}\lambda^{s-2r-1}e^{-(x+k)\lambda}d\lambda \longrightarrow (k+x)^{2r-s}\gamma(s-2r,\lambda_0(k+x))$, and hence (4.5) tends to zero for $\lambda \in (0,\lambda_0)$. For the case (ii) when $\lambda \in (\lambda_0,\infty)$ after intergrating both sides of (4.2) with respect to the inproper $\xi(\lambda) = \lambda^{-2r-1}\exp(-k\lambda); \lambda \in (\lambda_0,\infty)$, we have

$$\begin{array}{l} \int_{\lambda_0}^b\int_0^\infty \left(\left(\frac{\widetilde{u}}{\widehat{u}}-\ln\frac{\widetilde{u}}{u}-1\right)f(x,\lambda)\xi(\lambda)\,dxd\lambda\right. \\ \leq \int_{\lambda_0}^b\int_0^\infty \left(\frac{\widetilde{u}}{\widehat{u}}-\frac{\widetilde{u}}{\lambda^r}+\frac{\widehat{u}}{\lambda^r}-1\right)f(x,\lambda)\xi(\lambda)\,dxd\lambda \end{array} \tag{4.6}$$

After interchanging the order of integration and substituting for $\widehat{u}$, the right hand side of (4.6) becomes

$$\begin{array}{l} \int_0^\infty \frac{\widetilde{u\Gamma(s-2r,\lambda_0(k+x))}}{\Gamma(s-r,\lambda_0(k+x))\Gamma(s)}(x+k)^r x^{s-1}\int_{\lambda_0}^b\lambda^{s-r-1}e^{-(x+k)\lambda}d\lambda dx \\ -\int_0^\infty \frac{\widetilde{ux^{s-1}}}{\Gamma(s)}\int_{\lambda_0}^b\lambda^{s-2r-1}e^{-(x+k)\lambda}d\lambda dx \\ +\int_0^\infty \frac{\Gamma(s-r,\lambda_0(k+x))}{\Gamma(s-2r,\lambda_0(k+x))\Gamma(s)}(x+k)^{-r}x^{s-1}\int_{\lambda_0}^b\lambda^{s-2r-1}e^{-(x+k)\lambda}d\lambda dx \\ -\int_0^\infty x^{s-1}\int_{\lambda_0}^b\lambda^{s-r-1}e^{-(x+k)\lambda}d\lambda dx. \end{array} \tag{4.7}$$

Now, as $b \longrightarrow \infty, \int_{\lambda_0}^b\lambda^{s-r-1}e^{-(x+k)\lambda}d\lambda \longrightarrow (k+x)^{r-s}\Gamma(s-r,\lambda_0(k+x))$ and $\int_{\lambda_0}^b\lambda^{s-2r-1}e^{-(x+k)\lambda}d\lambda \longrightarrow (k+x)^{2r-s}\Gamma(s-2r,\lambda_0(k+x))$, and hence (4.7) goes to zero for $\lambda \in (\lambda_0,\infty)$. So in each cases

$$0 \leq \int_0^\infty \left(\frac{\widetilde{u}}{\widehat{u}}-\ln\frac{\widetilde{u}}{\widehat{u}}-1\right)f(x,\lambda)\,dx \leq 0$$

and this implies that

$$\int_0^\infty \left(\frac{\widetilde{u}}{\widehat{u}}-\ln\frac{\widetilde{u}}{\widehat{u}}-1\right)f(x,\lambda)\,dx = 0$$

i.e., $\widetilde{u} = \widehat{u}$ a.e., proving admissibility of $\widehat{u}(x)$.

The theorem can also be applied to the pareto distribution with density

$$f(y,\lambda,a) = \lambda a^\lambda y^{-\lambda-1} \qquad\qquad a < y < \infty, a, \lambda > 0$$

and

$$Y=\sum \lim_{i=1}^{n} \ln Y_i, s=n, \widehat{u}=\frac{\Gamma(n-r)}{\Gamma(n-2r)}\left(-n\ln a+\sum \lim_{i=1}^{n} \ln Y_i\right)^{-r}; r<\frac{n}{2},$$

the generalized Laplace distribution with density

$$f(y,b,k)=\frac{k}{2b\Gamma\left(\frac{1}{k}\right)}e^{-\frac{|y|^k}{b^k}} \qquad y\in\Re, b, k>0$$

and $\widehat{u}=\frac{\Gamma\left(\frac{n+r}{k}\right)}{\Gamma\left(\frac{n+2r}{k}\right)}\left(\sum \lim_{i=1}^{n}|Y_i|^k\right)^{r/k}$, where $r>-\frac{n}{2}$, the generalized gamma distribution with density

$$f(y,p,\lambda,a)=\frac{|\alpha|}{\Gamma\left(\frac{p}{\alpha}\right)}\lambda^{\frac{p}{\alpha}}y^{p-1}e^{-\lambda y^{\alpha}} \qquad 0<y<\infty, p\alpha>0$$

where p,α are given parameters, $Y=\sum \lim_{i=1}^{n} Y_i^{\alpha}, S=\frac{np}{\alpha}$.

We now consider the admissibility of the MRE estimators in the problem of Nile. The classical example of an ancillary statistic is known as the problem of Nile, originally formulated by Fisher (1959). Assume that X and Y are two positive valued random variable with joint density function

$$f(x,y,\tau)=e^{-\left(\tau x+\frac{1}{\tau}y\right)} \qquad ; x>0, y>0, \tau>0 \tag{4.8}$$

and that $(X_i,Y_i), i=1,\ldots,n$ is a random sample of n paired observations on (X,Y). Let $T=\sqrt{\frac{\overline{Y}}{\overline{X}}}, U=\sqrt{\overline{XY}}$, then (T,U) is a jointly sufficient, but not complete statistic for τ and U is ancillary. We use the loss (1.3) for constructing the minimum risk scale equivariant estimator of τ. Consider a nonrandomized rule $\delta(T,U)$ based on the sufficient statistic $\left(\overline{X},\overline{Y}\right)$ which is equivariant under the transformation

$$\begin{pmatrix} R \\ S \end{pmatrix}=\begin{pmatrix} c & 0 \\ 0 & \frac{1}{c} \end{pmatrix}\begin{pmatrix} \overline{X} \\ \overline{Y} \end{pmatrix} \qquad ; c>0.$$

We can see that all the scale equivariant estimator $\delta(T,U)$ must have the form $\delta(T,U)=T\phi(U)$, and so

$$\delta_{MRE}=\frac{k_0(2u)}{k_1(2u)}T \tag{4.9}$$

is the MRE estimator of τ, where $k_r(z)$ denotes the modified Bessel function of order r (Gradshteyn, I. S., and Ryzhik, I. M. (1980)) Since τ is a scale parameter, we consider the inverted Gamma as a prior, with density $\pi\lim_{\alpha,\lambda}(\tau)=\frac{\lambda^{\alpha}e^{-\lambda/\tau}}{\tau^{\alpha+1}\Gamma(\alpha)}; \tau>0, \lambda>0$, and the Bayes estimator of τ becomes

$$\delta_{Bayes}(\alpha,\lambda)=t\frac{\int_0^\infty \tau^{\alpha-1}e^{-\frac{\lambda\tau}{t}}e^{-u\left(\tau+\frac{1}{\tau}\right)}d\tau}{\int_0^\infty \tau^{\alpha}e^{-\lambda\frac{\tau}{t}}e^{-u\left(\tau+\frac{1}{\tau}\right)}d\tau}. \tag{4.10}$$

We can see that $\delta_{MRE}=\delta_{Bayes}(0,0)$, so δ_{MRE} is a generalized Bayes rule against the scale invariant improper priori $\pi(\tau)=\frac{1}{\tau};\tau>0$ and is therefore minimax. In the following theorem we show that it is admissible.

Theorem 4.2 : Let (X,Y) be distributed according to (4.8), and the minimum risk scal equivariant estimator δ_{MRE} is given by (4.9). Then δ_{MRE} is admissible for entropy loss function (1.3).

Pr **oof** : We use the method of Karlin (1985). Suppose that δ_{MRE} is not admissible, then there exists an estimator δ such tht the inequality

$$\begin{aligned}&\int_0^\infty\int_0^\infty\left(\tfrac{\delta(t,u)}{\tau}-\ln\tfrac{\delta(t,u)}{\tau}-1\right)g(t,u|\tau)\,dtdu\\&\qquad\le\int_0^\infty\int_0^\infty\left(\tfrac{\delta_{MRE}}{\tau}-\ln\tfrac{\delta_{MRE}}{\tau}-1\right)g(t,u|\tau)\,dtdu\end{aligned}$$

must be true for all τ and strict for at least one τ. The above inequality simplifies to

$$\begin{aligned}&\int_0^\infty\int_0^\infty\left(\tfrac{\delta(t,u)}{\delta_{MRE}}-\ln\tfrac{\delta(t,u)}{\delta_{MRE}}-1\right)g(t,u|\tau)\,dtdu\\&\qquad\le\int_0^\infty\int_0^\infty\left(\tfrac{\delta(t,u)}{\delta_{MRE}}-\tfrac{\delta_{MRE}}{\tau}+\tfrac{\delta_{MRE}}{\tau}-1\right)g(t,u|\tau)\,dtdu\end{aligned} \tag{4.11}$$

Now, let $\pi(\tau)=\frac{1}{\tau}$ and $0<a<b<\infty$. Intergrating both sides of (4.11) with respect to π we get

$$\begin{aligned}&\int_a^b\int_0^\infty\int_0^\infty\left(\tfrac{\delta(t,u)}{\delta_{MRE}}-\ln\tfrac{\delta(t,u)}{\delta_{MRE}}-1\right)g(t,u|\tau)\,\pi(\tau)\,dtdud\tau\\&\le\int_a^b\int_0^\infty\int_0^\infty\left(\tfrac{\delta(t,u)}{\delta_{MRE}}-\tfrac{\delta(t,u)}{\tau}+\tfrac{\delta_{MRE}}{\tau}-1\right)g(t,u|\tau)\,\pi(\tau)\,dtdud\tau\end{aligned} \tag{4.12}$$

Then, we simplinfying the right hand side of (4.12) , we have

$$\begin{aligned}&\int_a^b\int_0^\infty\int_0^\infty\tfrac{2\delta(t,u)k_1(2u)u^{2n-1}}{k_0(2u)n^{-2n}[(n-1)!]^2t^2}\tfrac{1}{\tau}e^{-nu\left(\frac{t}{\tau}+\frac{\tau}{t}\right)}dtdud\tau\\&-\int_a^b\int_0^\infty\int_0^\infty\tfrac{2\delta(t,u)u^{2n-1}}{n^{-2n}[(n-1)!]^2\tau t}\tfrac{1}{\tau}e^{-nu\left(\frac{t}{\tau}+\frac{\tau}{t}\right)}dtdud\tau\\&+\int_a^b\int_0^\infty\int_0^\infty\tfrac{2k_0(2u)u^{2n-1}}{k_1(2u)n^{-2n}[(n-1)!]^2\tau}\tfrac{1}{\tau}e^{-nu\left(\frac{t}{\tau}+\frac{\tau}{t}\right)}dtdud\tau\\&-\int_a^b\int_0^\infty\int_0^\infty\tfrac{2u^{2n-1}}{n^{-2n}[(n-1)!]^2t}\tfrac{1}{\tau}e^{-nu\left(\frac{t}{\tau}+\frac{\tau}{t}\right)}dtdud\tau.\end{aligned} \tag{4.13}$$

Now, interchanging the order of integration, (4.13) bocomes

$$\begin{array}{l} \int_0^\infty \int_0^\infty \frac{2\delta(t,u)k_1(2u)u^{2n-1}}{k_0(2u)n^{-2n}[(n-1)!]^2t^2} \int_a^b \tau^{-1}e^{-nu\left(\frac{t}{\tau}+\frac{\tau}{t}\right)}dtdud\tau \\ -\int_0^\infty \int_0^\infty \frac{2\delta(t,u)u^{2n-1}}{n^{-2n}[(n-1)!]^2t} \int_a^b \tau^{-2}e^{-nu\left(\frac{t}{\tau}+\frac{\tau}{t}\right)}dtdud\tau \\ +\int_0^\infty \int_0^\infty \frac{2k_0(2u)u^{2n-1}}{k_1(2u)n^{-2n}[(n-1)!]^2} \int_a^b \tau^{-2}e^{-nu\left(\frac{t}{\tau}+\frac{\tau}{t}\right)}dtdud\tau \\ -\int_0^\infty \int_0^\infty \frac{2u^{2n-1}}{n^{-2n}[(n-1)!]^2t} \int_a^b \tau^{-1}e^{-nu\left(\frac{t}{\tau}+\frac{\tau}{t}\right)}dtdud\tau. \end{array} \tag{4.14}$$

using the transformation $\frac{\tau}{t} = v$ and $nu \longrightarrow u$, (4.14) simplifies to

$$\begin{array}{l} \int_0^\infty \int_0^\infty \frac{2\delta(t,u)k_1(2u)u^{2n-1}}{k_0(2u)[(n-1)!]^2t^2} \int_{\frac{a}{t}}^{\frac{b}{t}} v^{-1}e^{-u\left(\frac{1}{\nu}+\nu\right)}d\nu dtdu \\ -\int_0^\infty \int_0^\infty \frac{2\delta(t,u)u^{2n-1}}{[(n-1)!]^2t^2} \int_{\frac{a}{t}}^{\frac{b}{t}} v^{-2}e^{-u\left(\frac{1}{\nu}+\nu\right)}d\nu dtdu \\ +\int_0^\infty \int_0^\infty \frac{2k_0(2u)u^{2n-1}}{k_1(2u)[(n-1)!]^2t} \int_{\frac{a}{t}}^{\frac{b}{t}} v^{-2}e^{-u\left(\frac{1}{\nu}+\nu\right)}d\nu dtdu \\ -\int_0^\infty \int_0^\infty \frac{2u^{2n-1}}{[(n-1)!]^2t} \int_{\frac{a}{t}}^{\frac{b}{t}} v^{-1}e^{-u\left(\frac{1}{\nu}+\nu\right)}d\nu dtdu. \end{array} \tag{4.15}$$

Now, letting $a \longrightarrow 0$ and $b \longrightarrow \infty$, we can see that $\int_{\frac{a}{t}}^{\frac{b}{t}} v^{-1}e^{-u\left(\frac{1}{v}+v\right)}dv \longrightarrow 2k_0(2u)$, and $\int_{\frac{a}{t}}^{\frac{b}{t}} v^{-2}e^{-u\left(\frac{1}{v}+v\right)}dv \longrightarrow 2k_{-1}(2u)$, and so (4.15) reduces to

$$\begin{array}{l} \int_0^\infty \int_0^\infty \frac{\delta(t,u)4u^{2n-1}}{t^2[(n-1)!]^2}\left(k_1(2u)-k_{-1}(2u)\right)dtdu \\ +\int_0^\infty \int_0^\infty \frac{4u^{2n-1}}{[(n-1)!]^2}\left(\frac{k_0(2u)k_{-1}(2u)}{k_1(2u)}-k_0(2u)\right)-dtdu \end{array} \tag{4.16}$$

using the recurrence relation $k_{v-1}(z) - k_{v+1}(z) = -\frac{2v}{z}k_v(z)$, (4.16) rdeuces to to zero, and hence

$$0 \leq \lim_{a\to 0, b\to\infty} \int_a^b \int_0^\infty \int_0^\infty \left(\frac{\delta(t,u)}{\delta_{MRE}} - \ln\frac{\delta(t,u)}{\delta_{MRE}} - 1\right) g(t,u|\tau)\,\pi(\tau)\,dtdud\tau \leq 0, \tag{4.17}$$

Note that range of integrations are Positive side of the real line. Also the integrand is strictly non-negative since $(y - \ln y - 1) \geq 0 \forall y \geq 0$. Then for (4.17) implies that

$$\frac{\delta(t,u)}{\delta_{MRE}} - \ln\frac{\delta(t,u)}{\delta_{MRE}} - 1 = 0 \qquad\qquad \forall t,u$$

and this happens if and olny if $\delta(t,u) = \delta_{MRE}$ since $y - \ln y - 1 \geq 0 \forall y \geq 0$ and minimum occures at $y = 1$, contradicting the earlier assumption that $\delta(T,U) \neq \delta_{MRE}$.
Note that when the loss is (1.5), the MRE estimator of τ is $\widehat{\tau}_{MRE} = T\phi_*(U)$, where $\phi_*(U)$ must satisfy the following equation

$$\begin{aligned}\phi_*^2(u)\left(\tfrac{u+\phi_*^{-1}(u)}{u+\phi_*(u)}\right)^{\frac{3}{2}} & k_1\left(2\sqrt{\left(u+\phi_*^{-1}(u)\right)(u+\phi_*(u))}\right)\\ &= k_{-1}\left(2\sqrt{\left(u+\phi_*^{-1}(u)\right)(u+\phi_*(u))}\right)\end{aligned} \tag{4.18}$$

and the unique Bayes estimator $\delta_{Bayes}(\alpha,\lambda)=\delta_B$ must satisfy the following equation

$$\begin{aligned}&k_{-\alpha-1}\left(2\sqrt{(\delta_B-ut^{-1})\left(\delta_B^{-1}-\lambda-ut\right)}\right)\\ &=\delta_B^2\left(\tfrac{\delta_B^{-1}-\lambda-ut}{\delta_B-ut^{-1}}\right)k_{-\alpha-1}\left(2\sqrt{(\delta_B-ut^{-1})\left(\delta_B^{-1}-\lambda-ut\right)}\right).\end{aligned} \tag{4.19}$$

Note that $\widehat{\tau}_{MRE}=\widehat{\tau}_{Bayes}(0,0)$, that is $\widehat{\tau}_{MRE}$ is a generalized Bayes rule against the scale invariant improper priori $\pi(\tau)=\frac{1}{\tau};\tau>0$, and is minimax. We conjecture that it is also admissible, but because of its complicated forms we don't have any proof.

Under the loss (1.6), the MRE estimator of τ is $\widehat{\tau}_{MRE}=T\phi^*(U)$, where $\phi^*(U)$ must satisfy the following integral equation

$$\int_0^\infty e^{\left(2a\phi^*(u)-u\right)t-\frac{u}{t}-e^{a(t\phi^*(u)-1)}}dt=e^a\int_0^\infty e^{\left(a\phi^*(u)-u\right)t-\frac{u}{t}-e^{a(t\phi^*(u)-1)}}dt \tag{4.20}$$

and the unique Bayes estimator δ_B under (1.6) must satisfies the following integral equation

$$\int_0^\infty \tau^{-\alpha}e^{(2a\delta_B-\frac{u}{t})\tau-(\lambda+ut)\frac{1}{\tau}-e^{a(\tau\delta_B-1)}}d\tau=e^a\int_0^\infty \tau^{-\alpha}e^{(a\delta_B-\frac{u}{t})\tau-(\lambda+ut)\frac{1}{\tau}-e^{a(\tau\delta_B-1)}}d\tau. \tag{4.21}$$

Note that $\widehat{\tau}_{MRE}=\widehat{\tau}_B$, whenever $\alpha\longrightarrow 0,\lambda\longrightarrow 0, i.e.\widehat{\tau}_{MRE}$ is a generalized Bayes rule with respect to scale invariant improper priori $\pi(\tau)=\frac{1}{\tau},\tau>0$, and therefore minimax, once again we cojectured that it is admissible, but becomes of its complicated form we don't have any proof.

5 Application to the Multivariate Case

Let $X_1,\ldots,X_n$ be i.i.d $N_\rho\left(\theta,\sum\right)$, where $\theta_{p\times 1}$ and $\sum \lim_{p\times p}$ are both unknown ($\theta\in\Re^p$, and $\sum$ p.d.) It is well known that $\left(\overline{X},S\right)$ is a complete sufficient statistic for $\left(\theta,\sum\right)$, where $\overline{X}=\frac{1}{N}\sum\lim_{i=1}^N X_i, S=\sum_{i=1}^N\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)'$. Let $X=\sqrt{N}\overline{X},\mu=\sqrt{N}\theta$ and $n=N-1$, then $X\sim N_p\left(\mu,\sum\right), S\sim W_p\left(\sum;n\right)$ and they are independentely distributed. We consider point estimation of covariance matrix $\sum$

and the precision matrix $\sum \lim^{-1}$, where $n > p+1$, when the loss is of the form (1.4), with $\Delta^* = tr\left(\widehat{\sum}\sum \lim^{-1} - I\right)$ for estimating $\sum$ and $\Delta_* = tr\left(\widehat{\sum}\sum \lim^{-1} - I\right)$ for estimating $\sum \lim^{-1}$, which is invariant and strictly Convex.

The tranformation $\left(\overline{X}, S\right) \longrightarrow \left(A\overline{X} + b, ASA^{'}\right)$ for $A_{p\times p}$ nonsingular and $b \in \Re^p$ is called an affine transformation, so as the loss is invariant under this group, an affine equivariant estimator turns out to be of the form cS for $\sum$ and dS^{-1} for $\sum^{-1}$, where c and d are positive constants.

It can be shown that the MRE of $\sum$ is c^*S, where $c^* = \frac{1}{2a}\left(1 - e^{\frac{-2ap}{np+2}}\right)$. For example for $\sum = diag\left(\sigma_{11}, \ldots, \sigma_{pp}\right)$, the MRE of $\sum$ under (1.4) with $\Delta = \Delta^*$ is c^*S^*, where $S^* = diag\left(S_{11}, \ldots, S_{pp}\right)$, and if $\sum = \sigma^2 I$, considering the estimators of the form $\widehat{\sum_s} = ctr\left(S\right) I$, then the best value of c uncdr (1.4) is $c_0 = \frac{c^*}{p}$. Since its riks is equal to the risk is equal to the risk c^*S, they are equivalent.

For estimating $\sum \lim^{-1}$ we look at the estimators of the form dS^{-1}. The following theorem shows that the optimum value of d doesn't have a closed form.

Theorem 5.1 : The optimal value of d, for estimators of the form dS^{-1} in estimation of $\sum \lim^{-1}$ under the loss (1.4) with $\Delta = \Delta_*$ must satisfy the following equation

$$e^{-ap}\frac{(ad)^{2np}\, 2^{\frac{-np}{2}}}{\Gamma_p\left(\frac{n}{2}\right)}\frac{\partial B_{2n}\left(\frac{-ap}{2}I\right)}{\partial d} - a\frac{p}{n-2} = 0 \tag{5.1}$$

where

$$B_\lambda\left(AZ\right) = \int \lim_{u>0} e^{tr(-AU)} e^{tr\left(-ZU^{-1}\right)} |A|^{-\lambda} |U|^{-\lambda - \frac{1}{2}(p+1)} dU$$
$$(A > 0, Z > 0, U > 0)$$

and $B_{-\lambda}\left(D\right) = B_\lambda\left(D\right)|D|^\lambda$, is the Bessel function of the secnod kind with matrix argument (see Herz 1955).
The MRE estimators obtained above are inadmissible and there are various way to improve over them.

(I) Improved Estimators Using S only: In this case we discuss L & U decomposition method, spectral decomposition method and Haff type estimators. James and Stein (1961) considered a somewhat smaller group of transtation, which result on the estimators of the form $\widehat{\sum} = L \bigtriangledown L^{'}$, where $S = LL^{'}, L$ is a lower triangular matrix, and ∇ is a diagonal matrix with positive diagonal elements. It can be shown using $S = LL^{'} \sim W_p\left(\sum, n\right), L_{ii}^2 \sim \chi^2_{(n-i+1)}, L_{ij}^2 \sim \chi^2_{(1)}; i > j$ that the best choice of the diagonal elements δ_i of $\bigtriangledown$ is $\delta_i^* = \frac{1}{2a}\left(1 - e^{-\frac{2ap}{np+2}}\right); i = 1, \ldots, p$ and hence

$\widehat{\sum}_L = L \bigtriangledown^* L^{'}$,is the MRE of $\sum$. If we consider the group of upper triangular matrices, with positive diagonal elements, then the best equivariant estimator of $\sum$ is $\widehat{\sum}_u = U \bigtriangledown U^{'}$, which is again the best multiple of S under (1.4) with $\Delta = \Delta^*$.

For estimating $\sum \lim^{-1}$, using the above method on S, the best upper triangular equivariant estimator $\widehat{\sum \lim_u^{-1}} = U^{-1} \nabla^{**} U^{'-1}$, where $\nabla^{**} = diag\left(\delta_1^*, \ldots, \delta_p^*\right)$. Unfortunately the calculations of δ_i^*'s are very complicated and in this case $\widehat{\sum \lim_u^{-1}}$ is not a multimple of S^{-1}.

Another method is spctral decomposition, motivated by Stein (1977b). He Considered the class of invariant orthogonally estimators $\sum$ of the form $\widehat{\sum} = R\Phi(L) R^{'}$, where $S = RLR^{'}$, R is the matrix of normalized eigenvectors $\left(RR^{'} = I = R^{'}R\right)$, $L = diag(L_1, \ldots, L_p)$ is the diagonal matrix of eigenvalues of S with $L_1 \geq L_2 \geq \ldots \geq L_p > 0$ and $\Phi(L) = diag(\Phi_1(L), \ldots, \Phi_p(L))$, $\phi_i(L) \geq 0$ are real valued functions $i = 1, \ldots, p$.

For estimating $\sum$ under (1.4) with $\Delta = \Delta^*$, we must choose $\phi_i(L)$'s such that to minimize the unbiased estimator of the risk function. We can show that $\widehat{\sum}$ dominates c^*S under (1.4) with $\Delta = \Delta^*$ if $p > 1$, and $a < 0$ where $\phi_i(L)$ is given by $\phi_i(L) = c^* L_i - \frac{L_i \log(L_i)\tau(u)}{b+u}; u = \sum \lim_{i=1}^{p} \log^2(L_i)$, and b is a positive constant, $\tau(u)$ is a strictly increasing in u and $E\left(\tau^{'}(u)\right) < \infty$. Haff (1980) introduced the following type of estimators for estimating $\sum, \widehat{\sum}_g = c^*S + g(S) I$, where $g(S) = g(u) = cut(u)$ and $u = \frac{1}{tr\left(S^{-1}\right)}$.

Now, we must find an appropriate form of $g(S)$ under the loss (1.4) with $\Delta = \Delta^*$ such that it dominates c^*S. We can show that for $a < 0$, $\widehat{\sum}_g$ dominates the MRE estimator of $\sum$ if

$$\begin{array}{ll} (i) & c > 0 \\ (ii) & t^{'}(u) < -\frac{M}{u}; M > 0 \\ (iii) & 0 < t(u) < \frac{2M}{p(n-p+1)} \end{array} \tag{5.2}$$

and for $a > 0$ the domintion conditions are

$$\begin{array}{ll} (i)^{'} & c < 0 \\ (ii)^{'} & t^{'}(u) > -\frac{M}{u}; M > 0 \\ (iii)^{'} & 0 < t(u) < \frac{2M}{p(n-p+1)} \end{array} \tag{5.3}$$

As an example we consider estimation of eigenvalues of $\sum$ i.e. $\lambda_1 \geq \ldots \geq \lambda_p > 0$. The usual estimator of $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ is $\widehat{\Lambda} = cL = cdiag(L_1, \ldots, L_p)$, where $L_1 \geq \ldots \geq L_p > 0$ are the eigenvalues of S, and c a suitable constant. Another class of estimators is $\widehat{\Lambda}_g = cL + g(u) I$, where $g(u)$ satisfies conditions (4.2) and (4.3). Then $\widehat{\Lambda}_g$ dominates $\widehat{\Lambda}$ with s_{ii} replaced by L_i. If

$$\sum = \sigma^2 \left[\rho 11^{'} - (\rho - 1) I\right] = \sigma^2 \begin{pmatrix} 1 \ \rho \ \ldots \ \rho \\ \rho \ 1 \ \ldots \ \rho \\ \vdots \\ \rho \ \rho \ \ldots \ 1 \end{pmatrix}, \text{where } -\frac{1}{p-1} < \rho < 1.$$

$$\Lambda = \sigma^2 \begin{pmatrix} 1+(p-1)\rho & 0 & \ldots & 0 \\ 0 & 1-\rho & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1-\rho \end{pmatrix} \text{ and } \quad S^* = \left[S^*_{ij}\right]_{p\times p} = A^{'} S A \text{ , where}$$

$S^*_{11} = \frac{tr(s)}{p}\left(1+(p-1)\, r\right), \sum \lim_{j=2}^{p} S^*_{jj} = \frac{tr(s)}{p}(1-r)(1-p)$ and $r = \frac{2\sum \lim_{u<u} S_{uv}}{(p-1)tr(S)}$. Notice that A is an orthogonal $p \times p$ matrix with first column constant and equal to $\frac{1}{\sqrt{p}}$. Therefore

$$\widehat{\Lambda} = c^* \begin{bmatrix} S^*_{11} & 0 & \cdots & \cdots & 0 \\ 0 & \frac{1}{p-1}\sum \lim_{j=2}^{p} S^*_{jj} & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{p-1}\sum \lim_{j=2}^{p} S^*_{jj} & \end{bmatrix} = c^* diag\left(m_1, \ldots, m_p\right)$$

is a risk equivalent estimator of $c^* S^*$, where $m_1 = S^*_{11}, m_i = \frac{1}{p-1}\sum \lim_{j=2}^{p} S^*_{ii}, i = 1, \ldots, p$.

(II) Improved Estimators Using S and $\overline{X}$: In this case instead of $\overline{X}$ we use the James-Stein estimator $\widehat{\theta}^{Js} = \left(1 - \frac{c^*_0}{T}\right)\overline{X}$, where $T = \overline{X}^{'} S^{-1} \overline{X}$ and $c^*_0 = \frac{p-2}{n(n-p+2)}$. So S becomes $S_* = \sum \lim_{1}^{N} \left(X_i - \widehat{\theta}^c\right)\left(X_i - \widehat{\theta}^c\right)^{'}$ in the bivariate case, where $\widehat{\theta}_c = \left(1 - \frac{c}{T}\right)\overline{X}$. We can see that $S_* = S + \frac{\text{constant}}{T^2}\left(\overline{X}\overline{X}^{'}\right)$. A new estimator of $\sum$ is a constant multiple of S_*. Another proposed esttimator of $\sum$ is $\widehat{\sum \lim_{c,\alpha}} = \widehat{\sum} + \frac{C}{(\overline{X}^{'} S^{-1} \overline{X})^{\alpha}} \overline{X}\overline{X}^{'}$, where $\widehat{\sum}$ is the best affine equivariant estimator of $\sum$.

Pal and Elfessi (1995) use this estimator, which is scale equivariant and uses both $\overline{X}$ and S. We can show that this estimator dominates $c^* S$ only if

$$c \geq \frac{e^{ap} - 1}{aN^{\alpha-1}} \cdot \frac{\Gamma\left(\frac{p}{2} + j + 1 - \alpha\right)}{\Gamma\left(\frac{p}{2} + j + 2(1-\alpha)\right)} \qquad\qquad j = 0, 1, \ldots.$$

or

$$c \geq \frac{e^{ap} - 1}{aN^{\alpha-1}} \tau(p, \alpha)$$

where

$$\tau(p,\alpha) = \max_j \frac{\Gamma\left(\frac{p}{2}+j+1-\alpha\right)}{\Gamma\left(\frac{p}{2}+j+2(1-\alpha)\right)}$$

where $\alpha \leq 1$, the above maximum is $\tau(p,\alpha) = \frac{\Gamma\left(\frac{p}{2}+1-\alpha\right)}{\Gamma\left(\frac{p}{2}+2(1-\alpha)\right)}$. The range of c in the case $\alpha = 1$ is $c \geq \frac{e^{ap}-1}{aN^{\alpha-1}}$, when $a > 0$ and $c > 0$. If $a > 0$ and $c < 0$ then $c \leq \frac{-e^{\frac{-2ap}{np+2}}}{aE\left\{\frac{\overline{X}' t \sum \lim^{-1} \overline{X}'}{T^{\alpha}}\right\}}$, when $\alpha = 1, c \leq \frac{-e^{\frac{-2ap}{np+2}}}{a(N-p)}$ if $a < 0$ and $c > 0$ for $\alpha = 1, c \geq \frac{-e^{\frac{-2ap}{np}}}{a(N-p)}$, and for other values of $\alpha, \widehat{\sum \lim}_{c,\alpha}$ is better than c^*S. A comparison of $\widehat{\sum \lim}_{c,\alpha}$ and $\widehat{\sum \lim}_g$ leads to the following theorem.

Theorem 5.2 : Under the loss function (1.4) with $\Delta = \Delta^*$, if $t(u)$ is an absolutely continuous and nonincreasing function, and

$$0 < t(u) \leq \frac{ac(N-p) + (1-2ac^*)\left(1-(1-2ac)^{-\frac{N-p}{2}}\right)}{a(N-p)c^*}$$

then for $a > 0$ the estimator $\widehat{\sum \lim}_{c,1}$ dominates $\widehat{\sum \lim}_g$, and for $a < 0$, the domination is reversed.

Remark 5.1: We can also estimate $tr\left(\sum\right)$ and $\det\left(\sum\right)$ under (1.4). Consider Bayesian and Empirical Bayesian estimation of $\sum$ and $\sum \lim^{-1}$, and find some improved estimators over $|\sum|$, along the methods of Brewster's sequential (1973) version.

Refrences

Akaike, H. (1977). On entropy maximization principle. In application of statistics (P.R. Krishnaiah, ed.) North - Holand, Amesterdam, 27-41.

Akaike, H. (1978). A new look at the Bayes procedure. Bimoetrika, 65, 53-59.

Brewster , J. F. (1973). A sequential version of Stein's variance estimator Technical Report, No. 45, University of Manitoba.

Brown, L. D. (1968). Inadmissibility of the usual estimators of Scale parameters in problem with unknown location and scale parameters. Ann-Math. Statist. 39,29-48.

Dey, D. K., Ghosh, M. & Srinivasan, C. (1987). Simultaneous estimation of parameters under entropy loss. J. Statist. Plann. Inference, 15, 347-363.

Fisher, R. A. (1959). Statistical Method, and scientific Inference. London, Oliver and Boyd.

Ghosh, Jk & Singh, R. (1970). Estimation of the reciprocal of the Scale Parameter of a gamma density. Ann. Inst. Statist. Math. 22, 51-55.

Ghosh, M. & Meeden, G. (1977). Admissibility of Linear estimators in the one parameter exponential family. Ann. statist. 5, 772-778.

Girshick, M. A. & Savage, L. J. (1951). Bayes and minimax estimates for quadratic loss functions, Proc. 2nd Berkely Symp. Math. Statist. Probab., 1, 53-73.

Gradshteyn. I., S, & Ryzhik, I. M. (1980). Table of intergrals series and Products. Sec. 3, 741, P. 340. Academic Press, Inc. New York.

Haff, L. R. (1982). Identities for the Wishart distribution with applications to regression. Sankhya Ser. B 44, 245-258.

Harris, T. J. (1992). Optimal controllers for nonsymmetric and non quadratic loss functions, Technometric, 34, No. 3, 298-306.

Herz, C. (1955). Bessel functions with matrix argument. Annals of Math., 61, 474-523.

James, W. & Stein, C. (1961). Estimation with quaratic loss. Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, 361-379, UNI of California Press.

Karlin, S. (1958). Admissibility for estimation with quadratic loss. Ann. Math. Statist., 29, 406-436.

Kuo, L. & Dey, D. K. (1990). On the admissibility of the linear estimators of the poisson mean using Linex loss functions. Statistics & Decisions, 8, 201-210.

Lehmann, E. L. (1951) . A general concept of unbiasedness. Ann. Math. Statist. 22, 578-592.

Lehmann, E. L. (1983). Theory of point estimation. New York, John Wiley and Sons.

Matta, J. M. & Cassella , G. (1990). Develpments in decision theoretic variance estimation. Statist. Sci. 5, 90-120.

Pal, N. & Elfessi, A. (1995). Improved estimation of a multivariate normal mean vector and the covariance matrix. How one affects the other.The Indian Journal of Statistics, 57, Series A, Pt, 2, 267-286.

Pitman, E. J. G. (1939). The estimation of the location and Scale parameters of a Continuous Population of any given form. Bimeterica, 30, 391-421.

Rukhin, A. L. & Ananda, M. M. A. (1992). Risk behavior of variance estimators in multivariate Normal distribution, Statistics and probability Letters, 13, 159-166.

Ralescu, D. & Ralescu, S. (1981). A class of nonlinear admissible estimators in the one parameter exponential family. Ann. Statist., 9, 177-183.

Schäbe, H. (1991). Bayes estimates under asymmetric loss. IEEE Transaction on Reliability, 19, 13-16.

Singh, R. (1972). Admissible estimators of λ^r in gamma distribution with quadatic

loss. Trabajos De Estadistica, 23, 129-134.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unkown mean, Ann. Inst. Statist. Math., 16, 155-160.

Varian. H. R. (1975) A Bayesian approach to real assessment. In: studies in Bayesian Econometric and statistics in Honor of Leonard J. Savage, eds. S. E. Fienberg and A. Zellner. North Holland, Amesterdam, 195-208.

Zellner, A. (1986) . Bayesian estimation and prediction using asymmetric loss function. J. Amer. Statist. Assoc. 81, 446-451.

Estimation of a Normal Mean Relative to Balanced Loss Functions

Sanjari Farsipour, N. and Asgharzadeh, A.

P12036

Department of Statistics, University Shiraz, Iran.

Abstract. Let $X_1, \cdots, X_n$ be a random sample from a normal distribution with mean θ and variance σ^2. The problem is to estimate θ with Zellner's (1994) balanced loss function, $L_B(\hat{\theta}, \theta) = \frac{\omega}{n}\sum_1^n (X_i - \hat{\theta})^2 + (1-\omega)(\theta - \hat{\theta})^2$, where $0 < \omega < 1$. It is shown that the sample mean $\overline{X}$, is admissible. More generally, we investigate the admissibility of estimators of the form $a\overline{X} + b$ under $L_B(\hat{\theta}, \theta)$. We also consider the weighted balanced loss function, $L_W(\hat{\theta}, \theta) = \omega q(\theta)\frac{\sum_1^n (X_i - \hat{\theta})^2}{n} + (1-\omega)q(\theta)(\theta - \hat{\theta})^2$, where $q(\theta)$ is any positive function of θ, and the class of admissible linear estimators is obtained under such loss with $q(\theta) = e^{\theta}$.

Keywords. Admissibility, Balanced Loss Function, Bayes Estimtor, Inadmissibility, Weighted Balanced Loss Function.

1 Introduction

Let $X_1, \cdots, X_n$ be a random sample from a $N(\theta, \sigma^2)$, with σ^2 known. This paper considers estimation of θ under the balanced loss function (BLF)

$$\begin{aligned} L_B(\hat{\theta}, \theta) &= \frac{\omega}{n}\sum_1^n (X_i - E_{\hat{\theta}}(X))^2 + (1-\omega)(\theta - E_{\hat{\theta}}(X))^2 \\ &= \frac{\omega}{n}\sum_1^n (X_i - \hat{\theta})^2 + (1-\omega)(\theta - \hat{\theta})^2, \end{aligned} \tag{1.1}$$

where $0 < \omega < 1$ and $\hat{\theta}$ is an estimator of θ. This loss function, introduced by Zellner (1994), is formulated to reflect two criteria, namely goodness of fit and precision of estimation. In the past, loss functions reflecting one or the other of these creteria, but not both, have been employed in analyses of means, regression and other estimation problems. For example, least squares estimation reflects goodness of fit considerations whereas use of quadratic loss functions involves a sole emphasis on precision of estimation. As is well known, sole emphasis on a precision of estimation criterion, for example mean squared error can often lead to biased estimators. In some circumstances bias is not important but in others, it is critical. On the other hand, use of a goodness of fit criterion leads to an estimate which gives good fit and is an unbiased estimator; however it may not be as precise as an estimator which is biased. Thus there is a need to provide a framework which combines goodness of fit, or lack of bias, and precision of estimation formally. The BLF framework meets this need. As mentioned above, first term of the r.h.s of (1.1) represents goodness of fit while the second term represents the precision of estimation. For a full discussion of properties of the BLF see Zellner (1994). For estimation under the BLF

and also for some references in this regard see Chung and Kim (1997), Chung, Kim and Song (1998) and Dey, Ghosh and Strawderman (1999) .

A generalization of the BLF (1.1), which is of interest, is

$$L_W(\hat{\theta},\theta) = \omega q(\theta)\frac{\sum_1^n (X_i-\hat{\theta})^2}{n} + (1-\omega)q(\theta)(\theta-\hat{\theta})^2, \tag{1.2}$$

where $0 < \omega < 1$ and $q(\theta)$ is any positive function of θ, which is called the weight function. This loss is called weighted balanced loss function (WBLF). It generalizes the BLF in sense that taking $q(\theta) = 1$.

The problem of admissibility of a class of linear estimators of the form $a\overline{X} + b$ in estimating a normal mean under a squared error loss has been studied by Karlin (1958) and Gupta (1966).
In this paper, we consider estimation of θ under the loss (1.1) and (1.2). In Section 2, we obtain a Bayes estimator of θ relative to the loss (1.1) and study the inadmissibility and admissibility of the estimators of the form $a\overline{X}+b$. In Section 3, the Bayes estimators relative to the WBLF (1.2) are discussed and the region of the inadmissibility and admissibility of the class of estimators of the form $a\overline{X}+b$ are derived under the WBLF (1.2) with $q(\theta) = e^{\theta}$.

2 Estimation of Mean under BLF

2.1 Bayes Estimators

Consider the Bayes estimator when the prior distribution on θ is normal with mean μ and variance τ^2. The posterior distribution is then normal with mean and variance given by

$$m = \frac{\frac{n\overline{X}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \quad \text{and} \quad \nu = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}},$$

respectively. The posterior risk of an estimator of $\hat{\theta}$ under the BLF is

$$\begin{aligned} E[\ L_B(\hat{\theta},\theta)|X] &= \frac{\omega}{n}\sum_1^n (X_i-\hat{\theta})^2 + (1-\omega)E[(\theta-\hat{\theta})^2|X] \\ &= \frac{\omega}{n}\sum_1^n (X_i-\hat{\theta})^2 + (1-\omega)[(m-\hat{\theta})^2+\nu] \end{aligned}$$

where $X = (X_1, ..., X_n)$.
To obtain the Bayes estimator of θ, it is enough to find an estimator $\hat{\theta}$ which minimizes $E[\ L_B(\hat{\theta},\theta)|X]$. Solving the equation $\frac{\partial E[\ L_B(\hat{\theta},\theta)|X]}{\partial\hat{\theta}} = 0$, we conclude that the Bayes estimator of θ under the loss (1.1) is

$$\begin{aligned} \hat{\theta}_B &= \omega\overline{X} + (1-\omega)m \\ &= \frac{n\tau^2+\sigma^2\omega}{n\tau^2+\sigma^2}\overline{X} + (1-\omega)\frac{\mu\sigma^2}{n\tau^2+\sigma^2}. \end{aligned} \tag{2.1}$$

The risk of $\hat{\theta}_B$ may be derived directly or deduced from Proposition 2.1.1 below where, for later use, we also give the risk function and Bayes risk of the linear

estimator $a\overline{X}+b$. The derivation is straightforward and is omitted.

Proposition 2.1.1. The risk function of the estimator $a\overline{X}+b$, relative to the BLF (1.1) is

$$R(\theta, a\overline{X}+b) = [(a-1)\theta+b]^2 + \frac{\sigma^2}{n}[(a-\omega)^2+\omega(n-\omega)], \tag{2.2}$$

and the Bayes risk of $a\overline{X}+b$, relative to the normal prior is

$$r(\pi, a\overline{X}+b) = \{(a-1)^2\tau^2+[(a-1)\mu+b]^2\} + \frac{\sigma^2}{n}[(a-\omega)^2+\omega(n-\omega)]. \tag{2.3}$$

2.2 Inadmissibility

In this section, the class of inadmissible linear estimators of the form $a\overline{X}+b$ is obtained. We shall now prove an inadmissibility result for linear estimators $a\overline{X}+b$, which is quite general and in particular does not require the assumption of normality.

Theorem 2.2.1. The estimator $a\overline{X}+b$ is inadmissible under the loss function (1.1) whenever one of the following conditions hold:
(i) $a>1$,
(ii) $a<\omega$,
(iii) $a=1$ and $b\neq 0$.

Proof:(i) If $a>1$, then $(a-\omega)^2>(1-\omega)^2$ and hence from (2.2)

$$\begin{aligned} R(\theta, a\overline{X}+b) &\geq \frac{\sigma^2}{n}[(a-\omega)^2+\omega(n-\omega)] \\ &> \frac{\sigma^2}{n}[(1-\omega)^2+\omega(n-\omega)] \\ &= R(\theta, \overline{X}). \end{aligned}$$

Thus, $a\overline{X}+b$ is dominated by $\overline{X}$.
(ii) If $a<\omega$, then $(a-1)^2>(\omega-1)^2$ and hence

$$\begin{aligned} R(\theta, a\overline{X}+b) &= [(a-1)\theta+b]^2 + \frac{\sigma^2}{n}[(a-\omega)^2+\omega(n-\omega)] \\ &= (a-1)^2[\theta+\frac{b}{a-1}]^2 + \frac{\sigma^2}{n}[(a-\omega)^2+\omega(n-\omega)] \\ &> (\omega-1)^2[\theta+\frac{b}{a-1}]^2 + \frac{\sigma^2}{n}[(a-\omega)^2+\omega(n-\omega)] \\ &\geq (\omega-1)^2[\theta+\frac{b}{a-1}]^2 + \frac{\sigma^2}{n}\omega(n-\omega)] \\ &= [(\omega-1)\theta+\frac{b(\omega-1)}{a-1}]^2 + \frac{\sigma^2}{n}\omega(n-\omega)] \\ &= R(\theta, \omega\overline{X}+\frac{b(\omega-1)}{a-1}) \ . \end{aligned}$$

Thus in this case, $a\overline{X}+b$ is dominated by $\omega\overline{X}+\frac{b(\omega-1)}{a-1}$.
(iii) When $a=1$, the risk function of $\overline{X}+b$ is

$$R(\theta,\overline{X}+b)=\frac{\sigma^2}{n}[(1-\omega)^2+\omega(n-\omega)]+b^2, \tag{2.4}$$

and the derivation of the risk in (2.4) with respect to b is $\frac{\partial}{\partial b}R(\theta,\overline{X}+b)=2b>0$, when $b>0$. Therefore, the risk in (2.4) is minimized at $b_0=0$. So, $R(\theta,\overline{X}+b)-R(\theta,\overline{X})=b^2>0$ for any real number $b\neq 0$. Thus $\overline{X}+b$ is dominated by $\overline{X}$ when condition (iii) holds.

Remark 2.2.1: Thus we see that in every case we are to look for admissible estimators of the form $a\overline{X}+b$ with (a,b) lying in the following strip of the $a-b$ plane:

$$\{(a,b)\ :\ \omega\leq a<1,\ \text{all } b\}\cup\{(1,0)\}\ .$$

2.3 Admissibility

In this section, admissible linear estimators are obtained. They are either proper Bayes estimators or generalized Bayes estimators relative to an appropriate limiting normal prior.

Theorem 2.3.1. The estimator $a\overline{X}+b$ is admissible under the BLF (1.1), whenever $\omega<a<1$.

Proof: From (2.1), we see that the coefficient $\frac{n\tau^2+\sigma^2\omega}{n\tau^2+\sigma^2}$ of $\overline{X}$ is strictly between ω and 1. Also since the loss (1.1) is strictly convex, (2.1) is the unique Bayes estimator and hence admissible. It follows that $a\overline{X}+b$ is admissible when $\omega<a<1$. □

It is seen that $\overline{X}$ is the limit of Bayes estimators (2.1) relative to the normal prior, when $\tau^2\longrightarrow\infty$, We proved the admissibility of $\overline{X}$ by the limiting Bayes method (due to Blyth (1951)).

Theorem 2.3.2. Under the BLF (1.1), $\overline{X}$ is admissible.
Proof: Suppose that $\overline{X}$ is not admissible, and without loss of generality, assume that $\sigma=1$. Then, there is an estimator δ^* such that

$$R(\theta,\delta^*)\leq R(\theta,\overline{X})=\frac{(1-\omega)^2+\omega(n-\omega)}{n}$$

for all θ , and with strict inequality for at least some θ. Now , $R(\theta,\delta)$ is a continuous function of θ for every δ so that there exists $\epsilon>0$ and $\theta_0<\theta_1$ such that

$$R(\theta,\delta^*)<\frac{(1-\omega)^2+\omega(n-\omega)}{n}-\epsilon$$

for all $\theta_0<\theta<\theta_1$. Let r_τ^* be the average risk of δ^* with respect to the prior distribution $N(0,\tau^2)$, and let r_τ be the Bayes risk of the Bayes estimator (2.1) with respect to $N(0,\tau^2)$. Then, by (2.3) it follows that

$$r_\tau=\frac{\tau^2(1-\omega)^2}{1+n\tau^2}+\frac{\omega(n-\omega)}{n}\ .$$

Hence

$$\frac{\frac{(1-\omega)^2+\omega(n-\omega)}{n} - r^*_\tau}{\frac{(1-\omega)^2+\omega(n-\omega)}{n} - r_\tau} = \frac{\frac{1}{\sqrt{2\pi}\tau}\int_{-\infty}^{+\infty}[\frac{(1-\omega)^2+\omega(1-\omega)}{n} - R(\theta,\delta^*)]e^{-\theta^2/2\tau^2}d\theta}{\frac{(1-\omega)^2}{n} - \frac{\tau^2(1-\omega)^2}{1+n\tau^2}}$$

$$\geq \frac{n(1+n\tau^2)\epsilon}{\tau(1-\omega)^2\sqrt{2\pi}}\int_{\theta_0}^{\theta_1} e^{-\theta^2/2\tau^2}d\theta \ .$$

The integrand converges monotonically to 1 as $\tau \longrightarrow \infty$ and hence by the Lebesgue monotone convergence theorem , the integral converges to $\theta_1 - \theta_0$ and hence the ratio converges to infinity. Thus , there exists $\tau_0 < \infty$ such that $r_{\tau_0^*} < r_{\tau_0}$, which contradicts the fact that r_{τ_0} is the Bayes risk for $N(0, \tau_0^2)$. It follows that $\overline{X}$ is admissible.

Remark 2.3.1 : $\overline{X}$ is minimax since it is admissible and it's risk is constant.

Remark 2.3.2: The case not covered yet is when $a = \omega$ and $b = 0$. It is seen that $\omega\bar{X}$ is the limit of Bayes estimators relative to the normal prior $N(0, \tau^2)$, when $\tau^2 \to 0$, and it is conjectured that it is admissible, but we do not have a proof. The problem is that the limiting Bayes argument does not work in this case.

3 Estimation of Mean under WBLF

In this section, we consider the Bayes estimator of θ under the weighted balanced loss function (1.2). The following proposition gives the general form of the Bayes estimator under the weighted balanced loss function.

Proposition 3.1. The Bayes estimator of θ, relative to the WBLF (1.2), is

$$\hat{\theta}_{WB} = \omega\hat{\theta_1} + (1-\omega)\hat{\theta_2} \ ,$$

where $\hat{\theta_1} = \overline{X}$ and $\hat{\theta_2} = E[\theta q(\theta)|X]/E[q(\theta)|X]$.

Remark 3.1: Note that $\hat{\theta}_{WB}$ is the value of θ that minimizes posterior expected loss. It is an average of $\bar{X}$, which minimizes the first term on the r.h.s. of (1.2) and $\hat{\theta_2}$, which minimizes the posterior expectation of the second term on the r.h.s. of (1.2). Accordingly, the Bayes estimator of θ can be expressed as linear combination of $\hat{\theta_1}$ and $\hat{\theta_2}$ by Proposition 3.1. □

Now, we consider the conjugate prior and calculate the Bayes estimators under several weight functions $q(\theta)$. To avoid complication in getting the Bayes estimators, we consider only two cases, $q(\theta) = \theta^2$ and $q(\theta) = e^\theta$. Suppose that prior of θ is $N(\mu, \tau^2)$. Then the posterior distribution $\pi(\theta|X)$ is $N(m, \nu)$, where m and ν given in section 2.1.

Case 1. $q(\theta) = \theta^2$:

Proposition 3.1 gives

$$\hat{\theta_2} = \frac{m^3 + 3m\nu}{m^2 + \nu} \quad \text{and} \quad \hat{\theta}_{WB} = \omega\bar{X} + \frac{(1-\omega)(m^3+3m\nu)}{m^2+\nu} \ .$$

Case 2. $q(\theta) = e^{\theta}$:
In this case, we obtain,

$$\hat{\theta_2} = m + \nu \quad \text{and} \quad \hat{\theta}_{WB} = \frac{n\tau^2 + \sigma^2\omega}{n\tau^2 + \sigma^2}\bar{X} + \frac{(1-\omega)(\sigma^2\tau^2 + \mu\sigma^2)}{n\tau^2 + \sigma^2}.$$

Note that, we have the above results replacing $E[\theta^3|X]$, $E[e^{\theta}|X]$ and $E[\theta e^{\theta}|X]$ with $m^3 + 3m\nu, e^{m+\nu/2}$ and $(m+\nu)e^{m+\nu/2}$ respectively.
Theorem 3.1. The estimator $a\bar{X} + b$ is admissible under the WBLF (1.2) with $q(\theta) = e^{\theta}$, whenever $\omega < a < 1$.

Proof. Suppose that the prior for θ is $N(\mu, \tau^2)$. Then the Bayes estimator of θ is

$$\hat{\theta}_{WB} = \frac{n\tau^2 + \sigma^2\omega}{n\tau^2 + \sigma^2}\bar{X} + \frac{(1-\omega)(\sigma^2\tau^2 + \mu\sigma^2)}{n\tau^2 + \sigma^2}. \tag{3.1}$$

Since the coefficient $\frac{n\tau^2+\sigma^2\omega}{n\tau^2+\sigma^2}$ of $\bar{X}$ is strictly between ω and 1. By the convexity of the WBLF (1.2), $\hat{\theta}_{WB}$ is the unique Bayes estimator and hence admissible. It follows that $a\bar{X} + b$ is admissible when $\omega < a < 1$. □

Remark 3.2: The estimator $a\overline{X} + b$ is inadmissible under the loss function (1.2), whenever one of the following conditions hold:
(i) $a > 1$,
(ii) $a < \omega$,
(iii) $a = 1$ and $b \neq 0$.

Proof: To see this, note that the risk function of $a\bar{X} + b$, relative to the WBLF is

$$R(\theta, a\overline{X} + b) = \{[(a-1)\theta + b]^2 + \frac{\sigma^2}{n}[(a-\omega)^2 + \omega(n-\omega)]\}q(\theta).$$

Now, since $q(\theta)$ is any positive function of θ, we can obtain the sufficient conditions for inadmissibility of $a\bar{X} + b$, exactly similar to Theorem 2.2.1.

Blyth, C.R. (1951). *On Minimax Statistical Decision Procedures and their Admissibility.* Ann. Math. Statist. **22**, 22-42.

Chung, Y., Kim, C. (1997). *Simultaneous Estimation of the Multivariate Normal Mean under Balanced loss function.* Commun. Statist - Theory Meth. **26**, 1599-1611.

Chung, Y., Kim, C. and Song, S. (1998). *Linear Estimators of a Poisson Mean under Balanced Loss Functions.* Statistics & Decisions. **16**, 245-257.

Dey, D.K, Ghosh, M. and Strawderman, W. (1999). *On Estimation with Balanced Loss Functions.* Statistics & Probability Letters. **45**, 2, 97-101.

Gupta, M.K. (1966). *On the Admissibility of Linear Estimates for Estimating the Mean of Distribution of the one Parameter Exponential family.* Calc. Statist. Assoc. Bull. **15**, 14-19.

Karlin, S. (1958). *Admissibility for Estimation with Quadratic loss.* Ann. Math. Statist. **29**, 406-436.

Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation,* Springer Verlag, New York.

Zellner, A. (1994). *Bayesian and Non-Bayesian Estimation Using Balanced Loss Functions.* Statistical Decision Theory and Related Topics V, (J.O. Berger and S.S. Gupta Eds). New York: Springer-Verlag, 377-390.

Minimax Estimation of Bounded Scale Parameter Under Entropy Loss Function

Sanjari Farsipour, N. and Jahedi, A.

A11218

Department of Statistics, Shiraz University, Iran.

Abstract. This paper is concerned with minimax estimation of a scale paramete θ when θ is restricted to the interval $[a, b]$ for some known $0 < a < b < \infty$. The loss function is entropy loss. It is shown, under some regularity conditions on density of the observation, that there exists an $m^* > 0$ such that, when $(a/b) - 1 < m^*$, a unique minimax estimator of θ exists. This minimax estimator is Bayes with respect to a two-point prior concentrated on a, b and this prior is least favorable.

Keywords. Bayes Estimation, Bounded Scale Parameter, Least Favorable Prior Distribution.

1. Introduction

For the problem of estimating a bounded real parameter θ, several authors have obtained minimax estimators of θ for the case where θ is restricted to an interval [a,b]. The normal mean case was considered by Casella and Strawderman [5], as well as by Zinzius [10], for squared-error loss; by Bischoff and Fieger[2] for the loss function $|d-\theta|^p$ with $p \geq 2$ and by Bischoff, Fieger and Wulfert[4] for linex loss. The case of the location parameter of an exponential distribution with squared-error loss was treated by Eichenauer[6]. Further, Eichenauer-Herrmann and Fieger[8] consider the case of distribution with support $[\theta, \theta + 1]$; they use a convex loss function. The case of a general location parameter with the loss function $|d - \theta|^p, p \geq 1$ can be found in Eichenauer-Herrmann and Ickstadt[9]; Bischoff and Fieger [3] consider a general parameter with absolute error loss. Results for the scale parameter case were obtained by Eichenauer-Hermann and Fieger [7] for squared-error loss and by Bischoff [1] for $|d - \theta|^p, p \geq 2$.
In this paper the case of restriction of the parameter, to the interval $[\theta_0, (1+m)\theta_0]$, with the loss $L(\delta, \theta) = \frac{\theta}{\delta_m} - ln\frac{\theta}{\delta_m} - 1, \delta \in I_m$, for θ is treated. It is shown under regularity conditions on the distribution of observations, that, when m is small, there exits a unique minimax estimator, which is Bayes with respect to a prior that is cocentrated on the set $\{\theta_0, (1 + m)\theta_0\}$ and this prior is least favorable.

2. Main Results

In this section, we present our main result, a minimax estimator for the scale parameter of a scale-invariant distribution under the entropy loss function. Our result derives from another one which we first state in the following much more general setting than that of this paper. Suppose χ denotes the sample space with an associated σ -algebra, β of subsets. The underlying experiment will lead to the measurement of random observable, X, taking values in χ. we take the sampling distribution of X to be an element of a space,$P = \{p_\theta\}, \theta \in \Theta$, indexed by the parameter space Θ, a compact convex subset of a linear space; Θ has a Borel

$\sigma - algebra$ of subsets denoted by τ. we denote the boundary of Θ by $\partial\Theta$. Assume $p \ll \mu$ so that, for each Θ , p_θ has a derivative $f(.|\theta)$ with respect to μ, a sigma-finite measure on the measurable space (χ, β) .We deem the decision space D in our estimation to be identical to the parameter space.
A nonrandomized decision rule $\delta : \chi \to \theta$ will be called feasible if $p_\theta(\delta \in \Theta) = 1$ for all $\theta \in \Theta$. Under the assumptions of the next theorem we restrict our attention to nonrandomized decision rules. Assume that the loss $L(d, \theta)$ is bounded below, and without essential loss of generality it is nonnegative, so that the risk function associated with feasible δ , $R(\delta, \theta) = \int L(\delta(x), \theta) f(x|\theta) d\mu(x)$ exist. Our analysis relies on the following general result.

Theorem 2.1: Suppose

i) δ_λ is a Bayes rule with respect to a distribution λ on (Θ, τ) for which $\lambda(\partial\Theta) = 1$;
ii) $R(\delta, \theta) = K$, a constant for all $\theta \in \partial\Theta$;
iii $\theta \to R(\delta_\lambda, \theta)$ is strictly convex on Θ.
Then δ_λ is minimax rule and λ a least favorable distribution. Furthermore, if δ_λ is the unique Bayes rule, then δ_λ is the unique minimax rule.

Proof 2.1: The hypotheses imply that

$$sup\{R(\delta_\lambda, \theta)|\theta \in \Theta\} = K. \tag{2.1}$$

From (2.1) it follows that (see e.g. Lehmann [12,p. 249]) δ_λ is minimax (and unique minimax when it is unique Bayes) and prior λ is least favorable. Now let us turn to the problem of central interest in this paper for which $\chi = R^n$ and μ represents lebesgue measure on R^n. The observable, X , has probability density function $f(x|\theta) = \theta^{-n} f(\theta^{-1}x)$ with respect to μ. We take f to be known and $\theta \in \Theta = I_m = [\theta_0, (1 + m)\theta_0]$ for some known $\theta_0 > 0, m > 0$. The loss function is

$$L(\delta, \theta) = \frac{\theta}{\delta_m} - \ln(\frac{\theta}{\delta_m}) - 1 \ \ \forall\theta, \delta \in I_m \tag{2.2}$$

Feasible estimators δ must now satisfy

$$p_\theta(\delta(x) \in I_m) = 1, \forall\theta \in I_m \tag{2.3}$$

concerning the support , $S(\theta)$,of P_θ we consider two cases, namely:
1)the case where $S(\theta)$ is independent of θ;
2)the case where $S(\theta) = [\beta\theta, \theta]^n$ for some known $\beta \leq 0$.
Note that, in case 2 and for all $\theta_1 < \theta_2$ with $\theta_i \in I_m, i = 1, 2$

$$S(\theta_0) \subset S(\theta_1) \subset S(\theta_2) \subset S((1 + m)\theta_0) \tag{2.4}$$

so that,for all $\theta \in I_m, x \in S(\theta_0)$, then $x \in S(\theta_1)$ and then $x \in S(\theta_2)$ and then $x \in S((1 + m)\theta_0)$. To simplify the notation, we let $S_1 = S(\theta_0), S_2 = S((1 + m)\theta_0)$. Furthermore, for any Borel measurable set A, for all $x \in A$ will mean for almost all $x \in A$ with respect to Lebesgue measure on A. Throughout the paper we work under the following basic condition.

Condition A

i) $f(\theta^{-1}x)$ has, for all $x \in S_1$, two continuous derivatives with respect to θ for each

$\theta \geq \theta_0$;
ii) for each $m > 0$ there exist three measurable and $\mu - integrable$ functions $k_i(x) = k_{i,m}(x), i = 0, 1, 2$, such that for $i = 0, 1, 2$

$$|\frac{\partial^i}{\partial \theta^i} f(\frac{x}{\theta})| \leq k_i(x)$$

for all $x \in S_1$ and all $\theta \in I_m$. To state our main result with a slight variation in earlier notation, let $\delta_{m,\lambda}$ denote the Bayes estimator of θ with respect to the two-point prior

$$\lambda(\theta) = \begin{cases} \lambda & if \quad \theta = \theta_0 \\ 1 - \lambda \; if & \theta = (1+m)\theta_0 \end{cases} \tag{2.5}$$

where $\lambda \in [0, 1]$.The following result can now be stated.

Theorem 2.2: There exist an $m^* \in (0, \infty)$, and for each $m \in (0, m^*)$ a $\lambda_m \in (0, 1)$ such that $\delta_m = \delta_{m,\lambda}$ is the unique minimax estimator of θ. Further, the prior (2.5) with $\lambda = \lambda_m$ is least favorable.

Remark 2.1: Note that each non-feasible estimator is on Θ, dominated by a feasible one. This implies that a feasible estimator which is minimax among the feasible ones is minimax among all estimators.

Proof of Theorem 2.2: First note that, if condition A is satisfied, then it is satisfied with $f(\theta^{-1}x)$ replaced by $f(x|\theta) = f(\theta^{-1}x)\theta^{-n}$.
Our proof of Theorem 2.2 uses the following lemmas to show that the conditions hypothesized in Theorem 2.1 obtain.□

Lemma 2.1: For all $x \in S_2$, the Bayes estimator $\delta_{m,\lambda}(x)$ is, for $\lambda < 1$, given by

$$\begin{aligned} \delta_{m,\lambda}(x) &= (1+m)\theta_0 I(x \notin S_1) \\ &+ \theta_0 \frac{\lambda f(\theta^{-1}x) + \frac{1-\lambda}{(1+m)^{n-1}} f(((1+m)\theta_0)^{-1}x)}{\lambda f(\theta^{-1}x) + \frac{1-\lambda}{(1+m)^n} f(((1+m)\theta_0)^{-1}x)} I(x \in S_1) \end{aligned} \tag{2.6}$$

Proof: Using (2.4) it can easily be seen that, for all $x \in S_2$, the posterior, given X=x distribution of θ is given by

$$P(\theta = \theta_0 | X = x) =$$

$$\{\frac{\lambda f({\theta_0}^{-1}x)}{\theta_0^n} I(\theta = \theta_0) + (1-\lambda)\frac{f(((1+m)\theta_0)^{-1}x)}{[(1+m)\theta_0]^n} I(\theta = (1+m)\theta_0)\} C(x)$$

Then $C(x) = \frac{\theta_0^n}{\lambda f({\theta_0}^{-1}x) + \frac{1-\lambda}{(1+m)^n} f(((1+m)\theta_0)^{-1}x)}$

$$\begin{cases} P(\theta = \theta_0 | X = x) = \frac{\lambda f({\theta_0}^{-1}x)}{\lambda f({\theta_0}^{-1}x) + \frac{1-\lambda}{(1+m)^n} f(((1+m)\theta_0)^{-1}x)} I(x \in S_1) \\ P(\theta = (1+m)\theta_0 | X = x) = 1 - P(\theta = \theta_0 | X = x) \end{cases} \tag{2.7}$$

Suppose $I_1 = I(x \in S_1)$ and $I_2 = I(x \notin S_1)$ further $\delta_{m,\lambda}(x)$ minimizes

$$R(\theta) = E\{\frac{\theta}{\delta_{m,\lambda}} - \ln(\frac{\theta}{\delta_{m,\lambda}}) - 1|X = x\}$$

and

$$\frac{\partial R(\theta)}{\partial \delta} = -\frac{E(\theta|x)}{\delta^2_{m,\lambda}(x)} + \frac{1}{\delta_{m,\lambda}(x)} = 0$$

implies that

$$E(\theta|x) = \delta_{m,\lambda}(x).\square \tag{2.8}$$

Remark 2.2: Although $\delta_{m,1}$ is not unique, for simplicity we choose it to be consistent with equation (2.6) as

$$\delta_{m,1}(x) = \theta_0 I(x \in S_1) + (1+m)\theta_0 I(x \notin S_1).$$

Any other choice would work equally well in the proof of Theorem 2.2,but the argument would be much less elegant.

Lemma 2.2: For all $x \in S_2$, $\delta_{m,\lambda}(x)$ satisfies:

i) $\delta_{m,0}(x) = (1+m)\theta_0 I_1 + (1+m)\theta_0 I_2, \delta_{m,1}(x) = \theta_0 I(x \in S_1) + (1+m)\theta_0 I(x \notin S_1)$ for all $m \geq 0$;
ii) $\delta_{m,\lambda}(x)$ is continues in λ for each $m > 0$ (2.9)
iii) $\delta_{m,\lambda}(x)$ is, for $x \in S_1$, strictly decreasing in λ for each $m > 0$

Proof: From (2.6) one easily obtaines (i) and (ii). For (iii) note that

$$\delta_{m,\lambda}(x) = (1+m)\theta_0 I(x \notin S_1) + \theta_0 \frac{\lambda A + \frac{1-\lambda}{(1+m)^{n-1}} B}{\lambda A + \frac{1-\lambda}{(1+m)^n} B} I(x \in S_1)$$

where, $A = f({\theta_0}^{-1}x)$ and $B = f(((1+m)\theta_0)^{-1}x)$

$$\begin{aligned}
&\frac{\partial(\delta_{m,\lambda}(x))}{\partial\lambda} \\
&= \theta_0 \frac{\lambda A^2 + \frac{(1-\lambda)AB}{(1+m)^n} - \frac{(1-\lambda)B^2}{(1+m)^{2n-1}} - (\lambda A^2 + \frac{(1-\lambda)AB}{(1+m)^{n-1}} - \frac{\lambda AB}{(1+m)^n} - \frac{(1-\lambda)B^2}{(1+m)^{2n-1}})}{\{\lambda A + \frac{1-\lambda}{(1+m)^n} B\}^2} \\
&= -\frac{m}{m+1} . \frac{AB}{\{\lambda A + \frac{1-\lambda}{(1+m)^n} B\}^2} < 0
\end{aligned}$$

for $m.AB > 0$. $\square$

Lemma 2.3: For each $m > 0$ there exists a unique $\lambda_m \in (0,1)$ such that

$$R(\delta_m, \theta_0) = R(\delta_m, (1+m)\theta_0),$$

where $\delta_m = \delta_{m,\lambda_m}$ and

$$R(\delta, \theta) = E_\theta(\frac{\theta}{\delta_m} - \ln(\frac{\theta}{\delta_m}) - 1)$$

Proof: Consider

$$R(\delta_{m,\lambda}, \theta_0) = \int_{S_1} (\frac{\theta}{\delta_{m,\lambda}} - \ln(\frac{\theta}{\delta_{m,\lambda}}) - 1).\frac{1}{\theta_0^n} f({\theta_0}^{-1}x)dx$$

where (see Lemma 2.2) $\frac{\delta_{m,\lambda}}{\theta_0} - 1$ is , for $x \in S_1$
i) non-negative,
ii) continuous and strictly decreasing in λ for each $m > 0$
iii) upperbounded by m
So, $R(\delta_{m,\lambda}, \theta_0)$ is , for each $m > 0$, continues and strictly decreasing in λ

$$R(\delta_{m,0}, \theta_0) = \frac{1}{(1+m)} - ln(\frac{1}{(1+m)}) - 1 > 0$$

$$\delta_{m,0} = (1+m)\theta_0$$

$$\begin{aligned} R(\delta_{m,0}, \theta_0) &= E_\theta(\frac{\theta_0}{(1+m)\theta_0} - ln(\frac{\theta_0}{(1+m)\theta_0}) - 1) \\ &= \frac{1}{(1+m)} - ln(\frac{1}{(1+m)}) - 1 > 0 \end{aligned}$$

$$\begin{aligned} R(\delta_{m,1}, \theta_0) &= E_\theta[\frac{\theta_0}{\theta_0 I_1 + (1+m)\theta_0 I_2} - ln(\frac{\theta_0}{\theta_0 I_1 + (1+m)\theta_0 I_2}) - 1] \\ &= [\frac{1}{I_1 + (1+m)I_2} - ln(\frac{1}{I_1 + (1+m)I_2}) - 1]p_{\theta_0}(x \in s_1) \\ &= [\frac{1}{I_1} - ln(\frac{1}{I_1}) - 1]p_{\theta_0}(x \in s_1) = 0 \end{aligned}$$

$$R(\delta_{m,0}, (1+m)\theta_0) = E_\theta[\frac{(1+m)\theta_0}{(1+m)\theta_0} - ln(\frac{(1+m)\theta_0}{(1+m)\theta_0}) - 1] = 0.$$

In the same way it can be shown that $R(\delta_{m,\lambda}, (1+m)\theta_0)$, is for each $m > 0$, continuous and strictly increasing in λ with

$$\begin{aligned} R(\delta_{m,1}, (1+m)\theta_0) &= E_\theta[\frac{(1+m)\theta_0}{\theta_0 I_1 + (1+m)I_2\theta_0} - ln(\frac{(1+m)\theta_0}{\theta_0 I_1 + (1+m)I_2\theta_0}) - 1] \\ &= E_\theta[\frac{(1+m)}{I_1 + (1+m)I_2} - ln(\frac{(1+m)}{I_1 + (1+m)I_2}) - 1] \\ &= [\frac{(1+m)}{I_1 + (1+m)I_2} - ln(\frac{(1+m)}{I_1 + (1+m)I_2}) - 1].P_{(1+m)\theta_0}(x \in s_1) \\ &= [(1+m) - ln(1+m) - 1].P_{(1+m)\theta_0}(x \in s_1) > 0. \end{aligned}$$

Thus there exists, for each $m > 0$, a unique $\lambda_m \in (0,1)$ with $R(\delta_m, \theta_0) = R(\delta_m, (1+m)\theta_0)$

Lemma 2.4: For each $m \geq 0$ and each $\lambda \in [0,1]$, $\frac{\partial^2}{\partial\theta^2} R(\delta_{m,\lambda}, \theta)$ is continuous in θ for each $\theta \geq \theta_0$.

Proof: Fix a $\lambda \in [0,1]$. Note that the risk function of $\delta_{m,\lambda}$ is given by

$$R(\delta_{m,\lambda}, \theta) = \int_{S_1} [\frac{\theta}{\delta_{m,\lambda}} - \ln(\frac{\theta}{\delta_{m,\lambda}}) - 1].f(x|\theta)dx$$
$$= R^*(m,\lambda,\theta) + \{[\frac{(1+m)\theta_0}{\theta} - \ln(\frac{(1+m)\theta_0}{\theta} - 1]\} \quad (2.10)$$

where

$$R^* = \int_{S_1} \{[\frac{\theta}{\delta_{m,\lambda}} - \ln(\frac{\theta}{\delta_{m,\lambda}}) - 1] - [\frac{(1+m)\theta_0}{\theta} - \ln(\frac{(1+m)\theta_0}{\theta}) - 1]\}.f(x|\theta)dx. \quad (2.11)$$

So it is sufficient to show that the second derivative of $R^*(m,\lambda,\theta)$ is continuous in θ for each $m \geq 0$ and $\theta \geq \theta_0$. For this proof, note that, by condition A, the integrand

$$T(x,\theta) = \{[\frac{\theta}{\delta_{m,\lambda}} - \ln(\frac{\theta}{\delta_{m,\lambda}}) - 1] - [\frac{(1+m)\theta_0}{\theta} - \ln(\frac{(1+m)\theta_0}{\theta}) - 1]\}.f(x|\theta) \quad (2.12)$$

has, for all $x \in S_1$ and each $\theta \geq \theta_0$,two continuous derivatives with respect to θ. Further, for all $x \in S_1$ and each $\theta \geq \theta_0$

$\frac{\delta_{m,\lambda}}{\theta} \leq 1+m$ and $\frac{\theta}{\delta_{m,\lambda}} \geq \frac{1}{1+m} = 1 - \frac{m}{1+m}$ and $1 - \frac{\theta}{\delta_{m,\lambda}} < \frac{m}{1+m}$ and

$$|1 - \frac{\theta}{\delta_{m,\lambda}}| \leq max(\frac{m}{m+1}, 1) \quad (2.13)$$

we know that $L(\delta_{m,\lambda}, \theta) = \frac{\theta}{\delta_{m,\lambda}} - ln\frac{\theta}{\delta_{m,\lambda}} - 1, \delta \in I_m$, for θ is decreasing and $0 \leq L \leq \frac{1}{1+m} - ln\frac{1}{1+m} - 1$, then

$$|L| \leq max(0, \frac{1}{1+m} - ln\frac{1}{1+m} - 1).$$

Now fix an $m \geq 0$. Then, for any $m'm$ and $\theta \in I_{m'}$, condition A and (2.13) imply that each of the derivatives $\frac{\partial^i}{\partial\theta^i} T(x,\theta), i = 0,1,2$, is , for $x \in S_1$, bounded by a μ-integrable function which dose not depend upon θ. This implies that, for $\theta \in I_{m'}$,

$$\frac{\partial^2}{\partial\theta^2} R^*(m,\lambda,\theta) = \int_{S_1} \frac{\partial^2}{\partial\theta^2} T(x,\theta) d\mu(x) \quad (2.14)$$

and that $\frac{\partial^2}{\partial\theta^2} R^*(m,\lambda,\theta)$ is continues in θ.

Lemma 2.5: For all $x \in S_2, \delta_m(x)$ is continuous in m for each $m \geq 0$.

Proof 2.5: First consider the case when m=0 and note that $\lim_{m\to 0}\lambda_m$ dose not exist. In fact, for m=0 any $\lambda \in [0,1]$ satisfies (see Lemma 2.3)

$$R(\delta_{m,\lambda},\theta_0) = R(\delta_{m,\lambda},(1+m)\theta_0).$$

However, δ_0 is well-defined because for $x \in S_2$, $\delta_{0,\lambda}(x) = \theta_0$ for all $\lambda \in [0,1]$. Further, for $x \in S_2$

$$\theta_0 \le \delta_m(x) \le (1+m)\theta_0$$

so

$$\lim_{m\to 0}\delta_m(x) = \theta_0,$$

which shows that δ_m is continuous in m at m=0. Now consider the case where $m > 0$. From (2.6) and condition A(i) it follows that $\delta_m(x)$ is continuous in m for all $x \in S_2$ if λ_m is continuous in m. This continuity of λ_m can be proved as follows. By the proof of Lemma 2.3, λ_m is the unique solution (in λ) to

$$G_m(\lambda) = R(\delta_{m,\lambda},\theta_0) - R(\delta_{m,\lambda},(1+m)\theta_0) = 0,$$

where $G_m(\lambda)$ is, for each $m > 0$, continuous and strictly decreasing in λ for $\lambda \in [0,1]$ and

$$\begin{aligned} G_m(0) &= R(\delta_{m,0},\theta_0) - R(\delta_{m,0},(1+m)\theta_0) \\ &= \frac{1}{1+m} - ln\frac{1}{1+m} - 1 - 0 = \frac{1}{1+m} - ln\frac{1}{1+m} - 1 > 0 \end{aligned}$$

Since $0 \le \frac{1}{1+m} \le 1$ and $m > 0$, $G_m(0)$ is decreasing .

$$\begin{aligned} G_m(1) &= R(\delta_{m,1},\theta_0) - R(\delta_{m,1},(1+m)\theta_0) \\ &= 0 - [(1+m) - ln(1+m) - 1]P_{(1+m)\theta}(x \in S_1) < 0. \end{aligned}$$

Further, for each $\lambda \in [0,1]$, $G_m(\lambda)$ is continuous in m for each $m > 0$. To see this,first note that, by the proof of lemma 2.4,

$$G_m(\lambda) = R^*(\delta_{m,\lambda},\theta_0) - R^*(\delta_{m,\lambda},(1+m)\theta_0) + m\frac{\theta_0}{\theta} - ln(1+m),$$

where $R^*(\delta_{m,\lambda},\theta)$ is defined in (2.11). The result then follows from condition A(ii) and the fact that, for all $x \in S_1$ and each $\lambda \in [0,1]$
i) $\delta_{m,\lambda}(x)$ is continuous in m by condition A(i).

ii) $\delta_{m,\lambda}(x)$ is, for each $m_0 > 0$ and uniformly in x, bounded in m in a neighborhood of m_0. Now take an $\epsilon > 0$ and let (λ,λ') be such that $0 \le \lambda < \lambda_m < \lambda' \le 1$, $|\lambda - \lambda'| \le \epsilon$.
Then $G_m(\lambda') < 0 < G_m(\lambda)$. Now find, by the continuity of G_m in $m, \eta > 0$ such that, for all $\eta^* \in (0,\eta)$,

$$|G_{m+\eta^*}(\lambda) - G_m(\lambda)| < \frac{1}{2}G_m(\lambda)$$

$$|G_{m+\eta^*}(\lambda') - G_m(\lambda')| < \frac{1}{2}|G_m(\lambda')|.$$

Then $G_{m+\eta^*}(\lambda') < 0 < G_{m+\eta^*}(\lambda)$ for all $\eta^* \in (0,\eta)$. Which implies $\lambda < \lambda_{m+\eta^*} < \lambda'$ for all $\eta^* \in (0,\eta)$ and thus $|\lambda_{m+\eta^*} - \lambda_m| < \epsilon$ for all $\eta^* \in (0,\eta)$. This prove that,for each $m > 0, \lambda$ is continuous in m.

Lemma 2.6: For each $m \geq 0$ and each $\theta \geq \theta_0, \frac{\partial^2}{\partial\theta^2} R(\delta_m, \theta)$ is continuous in m and this continuity is, for each $M > 0$,uniform in θ for $\theta \in I_M$.

Proof: It needs to be shown (see (2.10)) that, for every $m \geq 0, \epsilon > 0$ suffices to

$$H(m,\theta) = |\frac{\partial^2}{\partial\theta^2} R^*(m, \lambda_m, \theta) - \frac{\partial^2}{\partial\theta^2} R^*(m', \lambda_m, \theta)| < \epsilon,$$

for $|m - m'| < \eta, \theta \in I_M$. In order to prove this, first note that, by (2.14) and condition A, $H(m,\theta)$ is, for $\theta \in I_M$, upperbounded by a finite sum of term of the form

$$\frac{1}{\theta_0^l} \int_{S_1} |\delta_m^j(x) - \delta_{m'}^j(x)| K_{i,M}(x) dx.$$

Plus a finite sum of terms of the form

$$|(1+m)^i - (1+m')^j| \int_{S_1} K_{j,M}(x) dx,$$

where $i \in \{0,1,2\}$, $j \in \{1,2\}$, $l \in \{1,2,3,4\}$. This implies that it is sufficient to show that for $i \in \{0,1,2\}$, $j \in \{1,2\}$, each $m > 0$ and each $M > 0$, $\int_{S_1} \delta_M^j K_{j,M}(x) dx$ is continuous in m. But this continuity follows from lemma 2.5, the fact that, for each $m_0 \geq 0, \delta_m(x)$ is (uniformly in x for $x \in S_1$) bounded in a neighborhood of m_0 and the fact that the $k_{i,M}$ are integrable.

Lemma 2.7: There exists $m^{**} > 0$ such that, for each $m \in (0, m^{**})$,

$$\frac{\partial^2}{\partial\theta^2} R(\delta_m, \theta) > 0$$

for each $\theta \in I_m$.

Proof: Start with (2.10) and note that for $\theta \geq \theta_0$

$$\begin{aligned} B &= \frac{\partial^2}{\partial\theta^2}\{[\frac{(1+m)\theta_0}{\theta} - \ln(\frac{(1+m)\theta_0}{\theta}) - 1]\} \\ &= \frac{\partial}{\partial\theta}\{-[\frac{(1+m)\theta_0}{\theta^2} + \frac{1}{\theta}]\} \\ &= \frac{2(1+m)\theta_0}{\theta^3} - \frac{1}{\theta^2} \end{aligned}$$

for $m = 0$ and $\theta \geq \theta_0$

$$\frac{\partial^2}{\partial\theta^2}\{[\frac{(1+m)\theta_0}{\theta} - \ln(\frac{(1+m)\theta_0}{\theta}) - 1]\} = \frac{-\theta_0}{\theta^2} + \frac{1}{\theta} = \frac{1}{\theta^2}(\frac{2\theta_0 - \theta}{\theta}).$$

For the second derivative of $R^*(m,\lambda_m,\theta)$, note that, by the proof of Lemma 2.4, this derivative can be taken under the integral sign. But the second derivative of the integrand, evaluated at $m=0$,$\delta_m=\theta_0$ and $(1+m)\theta_0$ is

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}R^*(\delta_m,\theta) &= \frac{\partial^2}{\partial\theta^2}\int_{S_1}\{[\frac{\theta}{\delta_{m,\lambda}}-\ln(\frac{\theta}{\delta_{m,\lambda}})-1]\\ &\quad -[\frac{(1+m)\theta_0}{\theta}-\ln(\frac{(1+m)\theta_0}{\theta})-1]\}.f(x|\theta)dx\\ &= \frac{\partial}{\partial\theta}\{[\frac{\theta}{\delta_{m,\lambda}}-\ln(\frac{\theta}{\delta_{m,\lambda}})-1]-[\frac{(1+m)\theta_0}{\theta}-\ln(\frac{(1+m)\theta_0}{\theta})-1]\}.\\ &= \frac{1}{\delta_{m,\lambda}}-\frac{1}{\theta}+\frac{(1+m)\theta_0}{\theta^2}+\frac{1}{\theta}\\ &= \frac{1}{\delta_{m,\lambda}}+\frac{(1+m)\theta_0}{\theta^2}\end{aligned}$$

for $m=0$ we know that $\delta_{m,\lambda}=\theta_0$ and $(1+m)\theta_0=\theta_0$ then we have $\frac{1}{\theta_0}+\frac{\theta_0}{\theta^2}>0$

$$\frac{\partial^2}{\partial\theta^2}R(\delta_m,\theta)|_{m=0}=\frac{\partial^2}{\partial\theta^2}R^*(\delta_m,\theta)+B>0 \;\; for\; all \;\; \theta\in[\theta_0,2\theta_0]. \tag{2.15}$$

Now let $\theta^*\in[\theta_0,2\theta_0]$. Then, by Lemma 2.4 and (2.15),

$$C=min\{\frac{\partial^2}{\partial\theta^2}R(\delta_m,\theta)|_{m=0}|\theta\in[\theta_0,\theta^*]\}>0$$

Using Lemma 2.6 it then follows that, for $\epsilon\in(0,C)$, there exits $\eta>0$, independent of $\theta\in[\theta_0,\theta^*]$, such that

$$|\frac{\partial^2}{\partial\theta^2}R(\delta_m,\theta)-\frac{\partial^2}{\partial\theta^2}R(\delta_m,\theta)|_{m=0}|<\epsilon \;\;,\;\; 0\le m<\eta \; and \; \theta\in[\theta_0,\theta^*]$$

or, equivalently, such that

$$\frac{\partial^2}{\partial\theta^2}R(\delta_m,\theta)>\frac{\partial^2}{\partial\theta^2}R(\delta_m,\theta)|_{m=0}-\epsilon\ge C-\epsilon>0 \;\;,\;\; 0\le m<\eta \;\; and \;\; \theta\in[\theta_0,\theta^*]$$

Taking $m^{**}=min(\eta,(\frac{\theta^*}{\theta}-1))$ then gives the desired result.

Proof of theorem 2.2: By Lemma 2.1 and remark after it, the Bayes estimator of δ_m is achieved so that give us the condition hypothesized in the part (i) of theorem 2.1 and by applying theorem 2.3 which state $R(\delta_m,\theta_0)=R(\delta_m,(1+m)\theta_0)$ shows that the risk function for various values of m is constant. By Lemma 2.7, the convexity of $R(\delta_\lambda,\theta)$ is provided, finally the Bayes estimator δ_m is the unique minimax estimator because δ_m is the unique Bayes estimator for $\lambda=\lambda_m$. □

References

Bischoff, W. (1992).Minimax and $\Gamma-minimax$ estimation for functions of the bounded parameter of a scale parameter family under "L_p-loss" .*Statist. Decisions*, **10**,45-61.

Bischoff, W. and Fieger, W. (1992). Minimax and $\Gamma - minimax$ estimators for bounded normal mean under the loss $L_p(\theta, d) = |\theta - d|^p$. *Metrika*, **39**, 185-197.

Bischoff, W. and Fieger, W. (1993). On least favourable two point priors and minimax estimators under absolute loss. *Metrika*, **40**, 283-298.

Bischoff, W. and Fieger, W. and Wulfert, S. (1995). Minimax and $\Gamma - minimax$ estimation of a bounded normal mean under linex loss. *Statist. Decisions*, **13**, 287-298.

Casella, G. and Strawderman, W. E. (1981). Estimating a bounded normal mean. *Ann. Statist*, **9**, 870-878.

Eichenauer, J (1986). Least favourable two point priors in estimating the bounded location parameter of a noncentral exponential distribution. *Statist. Decisions*, **4**, 389-392.

Eichenauer-Herrmann, J. and Fieger, W. (1989). Minimax estimation in scale parameter families when the parameter interval is bounded. *Statist. Decisions*, **7**, 363-376.

Eichenauer-Herrmann, J. and Fieger, W. (1992). Minimax estimation under convex loss when the parameter interval is bounded. *Metrika*, **39**, 27-43.

Eichenauer-Herrmann, J. and Ickstadt, K. (1992). Minimax estimators for a bounded location parameter. *Metrika*, **39**, 227-237.

Van Eeden, C. and Zidek, J. V. (1998). Minimax Estimation of a bounded scale parameter for scale invariant square-error loss. *Statistic and Decisions*, **17**, 1-30.

Bilateral Laplace Transforms Application in Location Family

Shams, S.

P11064

Department of Mathematics, Alzahra University, Iran.

Abstract. It is well known that if a distribution function H, belongs to a location family, with location parameter θ, then $H_\theta(x) = H^*(x-\theta)$.

In many problems there has been interest in characterization of H^*, and the question of this kind arises when some information about the marginal distribution of the random variable X exists.

Suppose X and Θ, are two univariate random variables with distribution functions F and G, respectively, and the conditional distribution function X given θ, $H(x|\theta)$, belongs to a location family. In this paper , for any pairs of distribution functions F and G , the necessary and sufficient conditions for existence of a distribution function H, such that it can be the conditional distribution function belongs to a location family, are given. Also for given F and G, satisfying the above sufficient conditions, by bilateral Laplace transform the exact form of H is obtained.

1 Introduction

Suppose X and Θ are two random variables and the conditional distribution of X given $\Theta = \theta$ is unknown. In many problems there has been interest in characterization of the conditional distribution, because the distribution of X which we can observe is unconditional on Θ and we are sampling from a mixture distribution. The mixture models are useful because they make some properties of the derived distribution more transparent . Marshal and Olkin (1988) showed that in proportional hazard models the conditional distribution can be derived by a unified method, with a somewhat different Shams and Noorbalochi (2001) showed that in accelerated models it can be derived by using mellin transforms.

In this paper, in location models, after considering the necessary and sufficient conditions for existing such conditional distribution , by using bilateral Laplace transforms the conditional distribution can be derived.

2 Mixture of Location Distributions

In this section bilateral Laplace transforms are used to construct multivariate distributions with given marginals.

Let H and G be two univariate distribution functions and assume H belongs to a location family and θ is the location parameter.

Since $H(x-\theta)$ is a distribution function, it follows that the mixture

$$F(x) = \int H(x-\theta)dG(\theta) \tag{1}$$

is a distribution function .

For any specified pair of distribution functions G and F , the necessary and sufficient condition for existing a distribution function H satisfying (1), is given below.

Definition 1. A function f is called *completely monotonic* (CM) in an interval (finite or infinite, of any kind) iff it has derivatives of all orders there satisfying the condition:

$$(-1)^n f^{(n)}(x) \geq 0 \tag{2}$$

for each $n \geq 0$ and each x in the domain of definition.

Definition 2. If the integral $f(s) = \int_0^\infty e^{-sx} dF(x)$ converges for $a < s < b$, then $f(s)$ is called *analytic* there.

Definition 3. A real function $f(x, y)$ which is continuous in the square $(a \leq x \leq b, a \leq y \leq b)$ is of *positive type* if for every real function $g(x)$ continuous in $(a \leq x \leq b)$

$$\int_a^b \int_a^b f(x, y) g(x) g(y) dx dy \geq 0$$

Lemma 1. A function f on $(0, \infty)$ is the Laplace transform of a distribution function F;

$$f(s) = \int_0^\infty e^{-sx} dF(x) \tag{3}$$

if and only if it is completely monotonic in $(0, \infty)$ and $f(0+) = 1$.

Proof. The "only if" part is immediate, since

$$f^{(n)}(s) = \int_0^\infty (-x)^n e^{-sx} dF(x)$$

Turning to the "if" part , first we prove that f has a convergent Taylor series. Let $0 < s_0 < s < u$, then by Taylor's theorem, with the remainder term in the integral form, we have

$$f(s) = \sum_{j=0}^{k-1} f^{(j)}(u)(s-u)^j/j! + \frac{(s-u)^k}{(k-1)!} \int_0^1 (1-t)^{k-1} f^{(k)}(u + (s-u)t) dt \tag{4}$$

Because of (2), the last term in (4) is positive and does not exceed

$$\frac{(s-u)^k}{(k-1)!} \int_0^1 (1-t)^{k-1} f^{(k)}(u + (s_0 - u)t) dt$$

If k is even , then $f^{(k)} \downarrow$ and $(s-u)^k \geq 0$, while if k is odd then $f^{(k)} \uparrow$ and $(s-u)^k \leq 0$. Now by (4) with s replaced by s_0 the last expression is equal to

$$(\frac{s-u}{s_0-u})^k [f(s_0) - \sum_{j=0}^{k-1} f^{(j)}(u)(s_0-u)^j/j!] \leq (\frac{s-u}{s_0-u})^k f(s_0)$$

where the inequality is trivial, since each term in the sum on the left is positive by (2). Therefore as $k \to \infty$, the remainder term in (4) tends to zero and the Taylor

series for $f(s)$ converges.
Now for each $n \geq 1$ define the discrete subdistribution function F_n by

$$F_n(x) = \sum_{j=1}^{[nx]} \frac{nj}{j!}(-1)^j f^{(j)}(n)$$

This is a subdistribution function, since for each $\epsilon > 0$

$$1 = f(0) \geq f(\epsilon) \geq \sum_{j=1}^{k-1} f^{(j)}(n)(\epsilon - n)^j / j!$$

Letting $\epsilon \to 0$ and $k \to \infty$, we see that $F_n(\infty) \leq \infty$. The Laplace transform of F_n for $s > 0$ is:
$\int_0^\infty e^{-sx} dF_n(x) = \sum_{j=0}^\infty e^{-s(j/n)} \frac{(-n)^j}{j!} f^{(j)}(n)$

$$\sum_{j=0}^{\infty} \frac{1}{j!}(n(1 - e^{-s/n}) - n)^j f^{(j)}(n) = f(n(1 - e^{-s/n}))$$

the last equation form the Taylor series. Letting $n \to \infty$, we obtain for the limit of the last term $f(s)$, since f is continuous at each s. It follows, that $\{F_n\}$ converge, say to F, and that the Laplace transform of F is f. hence $F(\infty) = f(0) = 1$, and F is a distribution function.

Lemma 2. (Widder (1988)) A necessary and sufficient condition that the function $f(s)$ can be bilateral Laplace transform of a distribution function F,

$$f(s) = \mathcal{B}_F(s) = \int_{-\infty}^{\infty} e^{-sx} dF(x)$$

where the integral converges for $a < s < b$, is that (i) $f(s)$ should be analytic there, the kernel $f(s+s^{'})$ should be of positive type in the square $(a < 2s < b, a < 2s^{'} < b)$, and (ii) $f(0) = 1$.

Theorem 3. Let F, G be two distribution functions having bilateral Laplace transforms $\mathcal{B}_F, \mathcal{B}_G$, respectively. If $\mathcal{B}_H(s) = \frac{\mathcal{B}_F(s)}{\mathcal{B}_G(s)}$ satisfies the sufficient conditions of Lemma 2, then there exists a distribution function H satisfying (1) and it is determined uniquely (a.e) by

$$H(x) = \mathcal{B}^{(-1)}[\frac{\mathcal{B}_F(s)}{\mathcal{B}_G(s)}; x] \tag{5}$$

Here by $\mathcal{B}$ and $\mathcal{B}^{(-1)}$ we denote

$$\mathcal{B}_F(s) = \int_{-\infty}^{\infty} e^{-sx} dF(x) \Longleftrightarrow \mathcal{B}^{(-1)}[\mathcal{B}_F(s); x] = F(x)$$

Proof. The integral transform of H in (1) is of the convolution type, and if bilateral Laplace transform of the F and G exist, we have the well-known result that the bilateral Laplace transform of a convolution $H * G$ is the product of the transform that is ;

$$\int_{-\infty}^{\infty} e^{-sx} dF(x) = \int_{-\infty}^{\infty} e^{-su} dH(u) \int_{-\infty}^{\infty} e^{-s\theta} dG(\theta)$$

hence

$$H(x) = \mathcal{B}^{(-1)}[\frac{\mathcal{B}_F(s)}{\mathcal{B}_G(s)}; x]$$

This problem is similar to the problem of finding unbiased estimator of a location parameter, which has been discussed in Zacks(1970). Here we need to find an unbiased estimator for a distribution function, and this is done be the above assumptions.

3 Total Positivity of Location Models

Suppose (1) holds, and define the probability density functions of F, H and G by f, h and g, respectively , hence

$$f(x) = \int h(x-\theta)g(\theta)d\sigma(\theta) \tag{6}$$

Definition 4. Let X_1, X_2 have joint probability density function $f(x_1, x_2)$, then $f(x_1, x_2)$ is *totally positive of order* 2, TP_2 or alternatively $TP2(X_1, X_2)$, if

$$det \begin{vmatrix} f(x_1, x_2) & f(x_1, x_2^{'}) \\ f(x_1^{'}, x_2) & f(x_1^{'}, x_2^{'}) \end{vmatrix} \geq 0 \tag{7}$$

for all $x_1 < x_1^{'}, x_2 < x_2^{'}$ in domain of X_1, X_2.

Total positivity of order r (TP_r) is defined similarly in terms of determinants of order $1, 2, \ldots, r$. A function is totally positive of order infinity (TP_∞) if it is totally positive of all finite orders.

If h and g are Borel-measurable and totally positive of order r ,then for all σ-finite measure σ ,

$$f(x, z) = \int h(x, y)g(y, z)d\sigma(y) \tag{8}$$

is totally positive of order r (see Karlin(1968)) .

The following Lemma gives the necessary and sufficient condition for $k_i (i = 1, \ldots, n)$ to be TP_2 .

Lemma 4. A necessary and sufficient condition for $h(x-\theta)$ in (6) to be TP_2 is that $-logh$ in convex.

Proof. For $h(x-\theta)$, TP_2 is equivalent to

$$\frac{h(x-\theta^{'})}{h(x-\theta)} \leq \frac{h(x^{'}-\theta^{'})}{h(x^{'}-\theta)} \tag{9}$$

or

$$logh(x^{'}-\theta) + logh(x-\theta^{'}) \leq logh(x-\theta) + logh(x^{'}-\theta^{'})$$

where $x \leq x^{'}$ and $\theta \leq \theta^{'}$. Since $x-\theta = t(x-\theta^{'}) + (1-t)(x^{'}-\theta)$ and $x^{'}-\theta^{'} = (1-t)(x-\theta^{'}) + t(x^{'}-\theta)$, where $t = (x^{'}-x)/(x^{'}-x+\theta^{'}-\theta)$, a sufficient condition for this to hold is that function $-logh$ is convex.

To see that this condition is also necessary let $a < b$ be any real numbers , and let $x-\theta^{'} = a, x^{'}-\theta = b$ and $x^{'}-\theta^{'} = x-\theta$. Then $x-\theta = \frac{1}{2}(x^{'}-\theta+x-\theta^{'}) = \frac{1}{2}(a+b)$, and TP_2 implies

$$\frac{1}{2}[logh(a) + logh(b)] \leq logh[\frac{1}{2}(a+b)] \tag{10}$$

and this implies that $-logh$ is convex.

A density h for which $-logh$ is convex is called *strongly unimodal*.

Theorem 5. Suppose (6) holds and h are strongly unimodal functions, then f is (TP_2) in each pair of arguments, with the other arguments fixed.

Proof. This is an immediate consequence of Lemma (4).

References

Karlin,S.,(1968), *Total Positivity(Vol .1).* Stanford University Press.

Marshal, A.W. and Olkin , I.(1988), *Families of Multivariate Distributions,* Journal of the American Statistical Association, 83 ,834-841.

Shams, S. Noorbalochi, S., (2001) , *Mellin Transform Application in Accelerated Failure-Time Models,* Pakistan Journal of Applied Sciences , 2, 76-78.

Widder (1972), *The Laplace Transforms,* Princeton University Press.

Zacks, Sh., (1970), *The Theory of Statistical Inference,* John Wiley and Sons, Inc.

Probabilistic Analysis of Some Sorting Algorithms

Smythe, R. T.

P27001

Oregon State University, USA.

Abstract. One of the most common manipulations of a set of numbers is to sort them in increasing (or decreasing) order of magnitude. Many algorithms have been proposed for this purpose. For most of these, analysis has centered on calculation of average-case and worst-case performance.

If we assume the initial data array is uniformly distributed over all possible permutations, a probabilistic analysis of some of these algorithms is revealing. Often the number of comparisons made by the algorithm, when centered and scaled appropriately, converges to a limiting distribution as the size of the array grows without bound. However, the limiting distribution may be non-Gaussian and asymmetric, and in some cases quite complicated.

We examine the asymptotic behavior of three algorithms (Insertion Sort, Shell Sort, and Quick Sort), which exhibit three differnt types of limiting distributions.

Keywords. Random Permutation, Central Limit Theorem, Asymptotic Distribution, Insertion Sort, Shell Sort, Quick Sort.

0. Introduction. There are many reasons why one may wish to sort a given list of numbers in increasing order of magnitude. In statistics, for example, many nonparametric procedures are based on the order statistics, determination of which requires the sorting of part or all of a list of numbers. Many algorithms have been developed for this purpose. The average-case analysis (and sometimes, much more!) for a number of these may be found in the legendary book of Knuth (1973). Here we assume a random permutation model for the data and analyze probabilistically the number of comparisons performed by the algorithms. The reader is referred to the book of Mahmoud (2000) for details on the model and many results on the probabilistic analysis of algorithms. We will concentrate on three algorithms: the basic Insertion Sort, a modification of Insertion Sort known as Shell Sort, and Quicksort, an efficient divide-and-conquer algorithm. Our goal will be to present the asymptotic distribution of the number of comparisons as the size of the data array approaches infinity. We will see that although the asymptotics for Insertion Sort are quite standard, those for Shell Sort and Quick Sort take us into the class of more complicated distributions for which much remains unknown. Our treatment of Quick Sort is fairly brief, as good accounts of this exist elsewhere.

We will assume that our data are n real numbers from a continuous probability distribution. Since we are interested only in order properties of the data, and order is preserved by the probability integral transform, we can and will assume that the probability distribution of the data is uniform on $(0, 1)$. The random permutation model assumes that the ranks of the data are equally likely to be any of the permutations of $\{1, 2, \ldots, n\}$, each occurring with probability $1/n!$.

1. Insertion Sort. Insertion Sort is a simple algorithm with low efficiency, but its

simplicity and ease of coding makes it useful for small and medium-sized files. In addition, it is a basic building block for the more efficient Shell Sort.

Insertion Sort adds data in stages to a sorted file. At the i^{th} stage, it inserts the i^{th} key k into a sorted data file of $i-1$ elements. There are several methods of searching the sorted file, including linear, binary, and Fibonacci search. The search procedure finds the correct location, inserts the new key, and adjusts the positions of the other keys. For example, we examine how *backward linear insertion sort* works to sort the list $\{5, 2, 3, 1, 6, 4\}$.

The key to this analysis is the concept of an *inversion* in a permutation. In the original list $\{x_1, \ldots, x_n\}$, if the sequential rank of x_i is j, x_i causes $i-j$ inversions, that is, $i-j$ larger keys precede x_i in the input list. (Thus, in our example, the key 1 causes 3 inversions.) When x_i is inserted, $i-j$ comparisons, plus one for the "stopper", are required. In our example, the number of comparisons required would be

$$(0+1)+(1+1)+(1+1)+(3+1)+(0+1)+(2+1)=13.$$

In general, if C_n denotes the number of comparisons made for a list of size n,

$$C_n = \sum_{i=1}^{n}(1+V_i) := n + Y_n,$$

where V_i is the number of inversions caused by x_i and Y_n is the total number of inversions. It is not difficult to write down $P(C_n = k)$, but for our purposes it suffices to note that

$$E(C_n) \;=\; n \;+\; E(_n) \;\approx\; n^2/4,$$
$$Var(C_n) \;:=\; s_n^2 = Var(Y_n) \approx n^3/36.$$

The V_i are independent under the random permutation model, and using the Lindeberg central limit theorem, one can show (Lent and Mahmoud, 1996) that for linear search strategies,

$$\frac{C_n \;-\; E(C_n)}{(s_n^2)^{1/2}} \longrightarrow^{D}\lim\; \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ denotes the standard normal distribution. Thus for backward linear search, it follows that

$$\frac{C_n \;- n^2/4}{n^{3/2}} \longrightarrow^{D}\lim\; \mathcal{N}(0, 1/36).$$

The Berry-Esseen theorem may be invoked to show that the rate of convergence to normality is $1/\sqrt{(}n)$.

Linear insertion sort thus has a fairly simple asymptotic analysis, but the running time of $O(n^2)$ renders it unsuitable for large data files.

2. Shell Sort. Shell Sort generalizes the method of sorting by insertion, and consists essentially of several stages of Insertion Sort. The algorithm was proposed by Shell (1959) and is developed in Knuth's (1973) book. Although not as efficient as Quick Sort, Shell Sort is easy to implement and provides a practical, low-overhead method for medium-sized files.

Shell Sort performs k stages of Insertion Sort, where $k \geq 2$. For large values of k, the average running time for n keys can be made as small as $O(nlog^2 n)$, and even for $k = 2$, an optimized choice can give a $O(n^{5/3})$ average running time.

In k-stage Shell Sort, an integer sequence of length k, decreasing to 1, is chosen to give a faster running time than Insertion Sort. Let the chosen sequence be $t_k, t_{k-1}, \ldots, t_1 = 1$. The first stage in sorting n keys sorts keys that are t_k positions apart in the list. This gives t_k subarrays, each of length at most $\lceil n/t_k \rceil$. We sort these by ordinary Insertion Sort. At the next stage, the algorithm sorts t_{k-1} subarrays of keys that are t_{k-1} positions apart, using Insertion Sort on each of the $\lceil n/t_{k-1} \rceil$ arrays. Continuing in this fashion, the last stage (where the increment is 1) executes ordinary Insertion Sort. There has been a good deal of research directed to "best" choices of the sequence $\{t_j\}$; Sedgewick (1996) reviews much of this work.

An example may help to see how Shell Sort works. The simplest version, $(2, 1)$-Shell Sort, sorts the array

$$3 \quad 2 \quad 6 \quad 5 \quad 9 \quad 8 \quad 1 \quad 4 \quad 7$$

in two stages. In the first stage, using increments of 2, both the subarray of odd indexes and the subarray of even indices are sorted by ordinary Insertion Sort. This gives an interleaved array

$$\text{sorted odd positions:} \quad 1 \quad 3 \quad 6 \quad 7 \quad 9$$

$$\text{sorted even positions:} \quad 2 \quad 4 \quad 5 \quad 8$$

which is 2-sorted; that is, starting at any point, taking every second key gives an increasing sequence. The final step uses Insertion Sort to sort this 2-sorted array.

The stage of Shell Sort that uses the increment t_j will be referred to as the t_j-stage of the algorithm. We use the notation $Z_{(j)}$ to denote the j^{th} order statistic among $Z_1, \ldots, Z_r$, where the value of r will be evident in context.

To see the advantages of Shell Sort over Insertion Sort, consider $(h, 1)$-Shell Sort, which sorts h subarrays of size at most $\lceil n/h \rceil$ by Insertion Sort, then combines the h lists in another pass of Insertion Sort. The stages of sorting the h lists have a total number of comparisons of order $h(n/h)^2 \approx n^2/h$. Sorting of an h-sorted list requires an average of $n^{3/2}h^{1/2}$ comparisons (Knuth, 1973); choosing $h(n)$ to minimize $n^2/h + n^{3/2}h^{1/2}$ gives $h(n) \approx n^{1/3}$ and an average number of comparisons of $O(n^{5/3})$, an improvement over the $O(n^2)$ of ordinary Insertion Sort.

The difficulty in the stochastic analysis of Shell Sort is that after the first stage the data are no longer random. In $(2, 1)$-Shell Sort, for example, the second stage involves sorting a 2-sorted list; many of the original inversions have been removed at this stage (but some new ones may be added).

2.1 Analysis of $(2, 1)$-Shell Sort.

The mean and variance of $(2, 1)$-Shell Sort were given by Knuth (1973). The asymptotic distribution of $(2, 1)$-Shell Sort was first determined by Louchard (1986). A different analysis that permits generalization to $(h, 1)$-Shell Sort was given by Smythe and Wellner (2001). We will give a brief sketch of this analysis.

Assume for simplicity that n is even and call the elements in odd positions X's and those in even positions Y's. Thus our initial raw array is

$$X_1, Y_1, X_2, Y_2, \ldots X_{n/2}, Y_{n/2}.$$

The 2-stage of the algorithm orders the X's among themselves and the Y's among themselves. Let S_n be the number of comparisons that $(2,1)$-Shell Sort makes to sort n random keys, and let C_n be the number of comparisons made by Insertion Sort to sort n keys. The 2-stage of $(2,1)$-Shell Sort performs two insertion sorts on the subarrays $X_1, \ldots, x_{n/2}$ and $Y_1, \ldots, Y_{n/2}$. This requires

$$C_{n/2} + \tilde{C}_{n/2}$$

comparisons, where $\tilde{C}_j = \lim^D C_j$, and for all feasible i and j, C_i is independent of $\tilde{C}_j$.

The 1-stage now requires additional comparisons to sort the 2-sorted list. When we are about to insert $Y_{(j)}$, we place it among

$$\{X_{(1)}, \ldots, X_{(j)}\} \cup \{Y_{(1)}, \ldots, Y_{(j-1)}\}.$$

The Y's were ordered in the 2-stage, so $Y_{(j)}$ has no inversions with $\{Y_1, \ldots, Y_{(j-1)}\}$. The so-called sentinel version of Insertion Sort makes

$$C(\Pi_n) \;=\; n \;+\; I(\Pi_n)$$

comparisons to sort a permutation with $I(\Pi_n))$ inversions. Let V_j be the number of inversions $Y_{(j)}$ makes with all the elements that precede it, that is

$$V_j = 1_{\{X_{(1)} > Y_{(j)}\}} + \cdots 1_{\{X_{(j)} > Y_{(j)}\}},$$

for $j = 1, \ldots, n/2$. In a similar manner, define W_j to be the number of inversions that $X_{(j)}$ makes with all the elements that precede it. The number of inversions after the 2-stage is thus

$$I_n = \sum_{j=1}^{n/2} V_j \;+\; \sum_{j=1}^{n/2} W_j,$$

and the 1-stage then requires $n \;+\; I_n$ comparisons.

The overall number of comparisons S_n of $(2,1)$-Shell Sort is therefore given by the convolution

$$S_n = C_{n/2} \oplus \tilde{C}_{n/2} \oplus n \;\oplus I_n. \tag{1}$$

We have already noted that $C_{n/2}$, and thus $\tilde{C}_{n/2}$ as well, have asymptotic Gaussian distributions. So we focus on the distribution of I_n. The next result is key to the analysis (see Smythe and Wellner (2001) for the proof):

LEMMA 1. *The total number of inversions after the 2-stage, I_n, has the representation*

$$I_n \;=\; \sum_{j=1}^{n/2} T_n^{(j)},$$

where

$$T_n^{(j)} \;=\; \Big|\sum_{i=1}^{n/2}[1_{\{Y_i<X_j\}} - 1_{\{X_i<X_j\}}\Big|.$$

The usefulness of Lemma 1 is that it allows us to work directly with the X's and Y's, instead of with the order statistics. We can express the form in Lemma 1 in terms of empirical distribution functions. Let $F_n(t)$ denote the empirical distribution function of n i.i.d. random variables Z_i, that is,

$$F_n(t) \;=\; \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i\le t\}}.$$

Then

$$T_n^{(j)} \;=\; \frac{n}{2}\Big|\hat{F}_{n/2}(X_j) - \big(\tilde{F}_{n/2}(X_j) - 2/n\big)\Big|$$

where $\hat{F}_{n/2}$ and $\tilde{F}_{n/2}$ are the empirical distribution functions of the Y's and X's, respectively. Thus we may express I_n as

$$I_n \;=\; \frac{n}{2}\sum_{j=1}^{n/2}|\hat{F}_{n/2}(X_j) - \tilde{F}_{n/2}(X_j)| + o_p(n)$$

(where $o_p(n)$ denotes a term smaller than order n in probability, resulting from the "extra" $2/n$ in the expression for $T_n^{(j)}$), so that

$$\frac{I_n}{(n/2)^{3/2}} \;=\; \int_0^1 \sqrt{n/2}|\hat{F}_{n/2}(x) - \tilde{F}_{n/2}(x)|d\tilde{F}_{n/2}(x) + o_p(1). \qquad (2)$$

The empirical process converges in the Skorohod topology on $D[0,1]$, the space of right-continuous functions with left limits on $[0,1]$, giving

$$\sqrt{n}(F_n(t) \;-\; t) \;\longrightarrow \overset{D}{\lim}\; B(t),$$

where $B(t)$ is the (standard) Brownian bridge. Thus

$$\sqrt{n/2}\;|\hat{F}_{n/2}(x) - \tilde{F}_{n/2}(x)| \;\longrightarrow \overset{D}{\lim}\; |B_1(x) - B_2(x)|, \qquad (3)$$

where $B_1(t)$ and $B_2(t)$ are independent Brownian bridges. The Brownian bridge is a Gaussian process, and $B_1 - B_2$ has the distribution of $\sqrt{2} * (a\ Brownian\ bridge)$, so

$$B_1(t) - B_2(t) \;= \overset{D}{\lim}\; \mathcal{N}(0, 2t(1-t)).$$

Also, the empirical measure $d\tilde{F}_{n/2}(x)$ converges a.s. to uniform measure on $[0,1]$, by the Glivenko-Cantelli theorem. Using a special construction to put versions of $\hat{F}_{m/2},\ \tilde{F}_{n/2},\ B_1$ and B_2 on a common probability space, passing to the limit in (2) gives the result of Louchard (1986):

THEOREM 1. *Let W_n be the number of comparisons made in the 1-stage of $(2,1)$-Shell Sort to sort n random keys, so that $W_n = n + I_n$. Then*

$$\frac{W_n}{(n/2)^{3/2}} \longrightarrow \overset{D}{\lim}\; \sqrt{2}\int_0^1 |B(t)|dt,$$

where $B(t)$ is a standard Brownian bridge.

An infinite series for the c.d.f of the limiting distribution of Theorem 1, involving Airy functions, is given by Johnson and Killeen (1983). From (2) it can be shown that the convergence in distribution in Theorem 1 entails the convergence of moments of $W_n/(n/2)^{3/2}$ to those of $\sqrt{2}]\int_0^1 |B(t)|dt$. These moments were derived by Shepp (1982).

COROLLARY 1 .
(a) *The average number of comparison made by the 1-stage of* $(2,1)$*-Shell Sort is asymptotically equivalent to* $(\sqrt{\pi}/8\sqrt{2})n^{3/2}$.
(b) *The variance of the number of comparisons is asymptotically equivalent to* $(7/240-\pi/128)n^3$.

2.2 Analysis of $(h,1)$-Shell Sort

The analysis of $(h,1)$-Shell Sort proceeds along similar lines to those of the previous section. The only significant difference is that now we have h data "types" $X^1,\ldots,X^h$ instead of just X and Y. Thus the number of inversions caused by $X^k_{(i)}$, where $1\le k\le h$ and $1\le i\le\lceil n/h\rceil$, has to take account of the inversions caused with all the other X^j, where $j\neq k$. Let I^h_n denote the total number of inversions and let $(B^1,\ldots,B^h)$ be a vector of independent Brownian bridges. Proceeding as before, and assuming for simplicity that h divides n, we arrive at

THEOREM 2. *Let W^h_n be the number of comparisons made in the 1-stage of $(h,1)$-Shell Sort to sort n random keys, so that $W^h_n = n + I^h_n$. Then*

$$\frac{W^h_n}{(n/h)^3/2} \longrightarrow^{D}\lim\ W^h := \sum_{r<s}\int_0^1 |B^r(t)-B^s(t)|dt,$$

where the sum extends over all pairs (r,s) with $1\le r<s\le h$.

As far as I know, the form of the distribution of W^h for $h>2$ is not known. Of course the summands that comprise W^h have known distributions, but the (i,j) and (r,s) summands will be correlated if the pairs (i,j) and (r,s) share an integer. The first moment, which was found by Knuth (1973), follows easily from the representation. We do not have a closed form for the second moment, but it can be evaluated numerically.

COROLLARY 2.
(a) *The average number of comparisons made by the 1-stage of $(h,1)$-Shell Sort is asymptotically equivalent to*

$$\left(\frac{n}{h}\right)^{3/2}\binom{h}{2}\frac{\sqrt{\pi}}{4}.$$

(b) The variance of the number of compariosns is asymptotically equivalent to

$$\left(\frac{n}{h}\right)^3\left\{\binom{h}{2}\left(\frac{7}{30}-\frac{\pi}{16}\right)+h(h-1)(h-2)\left(C-\frac{\pi}{16}\right)\right\},$$

where

$$C \;:=\; \mathbf{E}\Big\{\int_0^1 |B_1(t)-B_2(t)|dt\int_0^1 |B_1(t)-B_3(t)|dt\Big\}$$

and $B_1(t), B_2(t), B_3(t)$ are independent (standard) Brownian bridges. The value of C is numerically determined to equal 0.2051.

For $s<t$, $Cov(B(s),B(t))=s(1-t)$, so C can be evaluated as a double integral once it is known how to calculate $E(|X||Y|)$, where (X,Y) is bivariate normal with mean $(0,0)$, unit variance, and covariances ρ. (See Wellner and Smythe (2001) for several ways of doing this.)

2.3. $(3,2,1)$**-Shell Sort.**

If k is a divisor of h, $(h,k,1)$- Shell Sort can be analyzed by the results of the previous section, but this case is not of practical interest as it has poor worst-case behavior. The simplest case when h and k are relatively prime is $(3,2,1)$-Shell Sort, and we now give a brief analysis of this. Recent work of Janson and Knuth (1997) gives detailed results on the mean number of comparisons for $(h,k,1)$-Shell Sort; for the $(3,2,1)$ case, we are able to find the variance and the limiting distribution of the final stage. It turns out that the last stage, which amounts to sorting a list that is both 3-sorted and 2-sorted, makes an asymptotically normal number of comparisons. As an example, we apply $(3,2,1)$-Shell Sort to sort the array

$$3 \quad 12 \quad 6 \quad 10 \quad 5 \quad 9 \quad 8 \quad 1 \quad 11 \quad 4 \quad 7 \quad 2.$$

The first stage creates 3 sorted lists of length 4:

$$\begin{array}{cccc} 3 & 4 & 8 & 10 \\ 1 & 5 & 7 & 12 \\ 2 & 6 & 9 & 11. \end{array}$$

The second stage takes the resulting 3-sorted list,

$$3 \quad 1 \quad 2 \quad 4 \quad 5 \quad 6 \quad 8 \quad 7 \quad 9 \quad 10 \quad 12 \quad 11,$$

and creates two sorted lists of length 6:

$$\begin{array}{cccccc} 2 & 3 & 5 & 8 & 9 & 12 \\ 1 & 4 & 6 & 7 & 10 & 11. \end{array}$$

We now have a list that is both 3-sorted and 2-sorted:

$$2 \quad 1 \quad 3 \quad 4 \quad 5 \quad 6 \quad 8 \quad 7 \quad 9 \quad 10 \quad 12 \quad 11,$$

and the final stage sorts this list.

For the general case, assume for convenience that n is a multiple of 6. Prior to any sorting, we will denote our raw data as

$$X_1, Y_1, Z_1, X_2, Y_2, Z_2, \ldots, X_{n/3}, Y_{n/3}, Z_{n/3},$$

where the X's, Y's, and Z's may be taken to be mutually independent. As before, we let C_n denote the number of comparisons made by (linear) insertion sort to sort n random keys. The initial stage of $(3, 2, 1)$-Shell Sort makes three runs of insertion sort on the X, Y, and Z subarrays, requiring

$$C^1_{n/3} \ + \ C^2_{n/3} \ + \ C^3_{n/3}$$

comparisons, where C^k_j, $k = 1, 2, 3$ are identically distributed for each j and C^1, C^2, C^3 are independent. We then have the 3-sorted list

$$X_{(1)}, Y_{(1)}, Z_{(1)}, \ldots, X_{(n)}, Y_{(n)}, Z_{(n)}.$$

The next stage of the algorithm 2-sorts this list: we make two lists,

$$X_{(1)}, Z_{(1)}, Y_{(2)}, X_{(3)}, Z_{(3)}, Y_{(4)}, \ldots, \tag{4}$$

and

$$Y_{(1)}, X_{(2)}, Z_{(2)}, Y_{(3)}, X_{(4)}, Z_{(4)}, \ldots, \tag{5}$$

and we sort each of these lists. Each of these lists is 3-sorted; let $\hat{C}_n$ denote the number of comparisons made by insertion sort to sort a 3-sorted array of length n. The 2-stage of the algorithm thus requires

$$\hat{C}^1_{n/2} \ + \ \hat{C}^2_{n/2}$$

comparisons, where $\hat{C}^1_j =\lim^D \hat{C}^2_j$ and $\hat{C}^1, \hat{C}^2$ are independent. (The asymptotic behavior of $\hat{C}^i_j$ was found in Section 2.2, taking $h = 3$.) The final stage of the algorithm sorts the 2-sorted and 3-sorted list. Denote the keys in list (4) by $U_1, U_2, ...$ and those in (5) by $V_1, V_2, ...$ When we are about to insert $V_{(j)}$, we place it among

$$\{U_{(1)}, \ldots, U_{(j)}\} \cup \{V_{(1)}, \ldots, V_{(j-1)}\}.$$

Then the overall number of comparisons (using the sentinel version of Insertion Sort) is given by the sum

$$S_n = C^1_{n/3} + C^2_{n/3} + C^3_{n/3} + \hat{C}^1_{n/2} + \hat{C}^2_{n/2} + n + I_n,$$

where all terms are independent, and I_n represents the number of inversions in the list that is both 2-sorted and 3-sorted. From our previous discussion, only the limiting distribution of I_n is as yet unknown.

Regard the X's, Y's, and Z's as three independent sets of $n/3$ i.i.d. observations, uniformly distributed on $(0, 1)$, giving rise to a set of n points in $(0, 1)$. Associate to each of these points a triple giving the parity of the numbers of X's, Y's, and Z's, respectively, that precede the point; for example, the triple OEO means that the number of X-predecessors is odd, the number of Y-predecessors is even, and the number of Z-predecessors is odd. Let

$$N_4 := \text{number of points of types } OEO, EOE.$$

The next lemma is key to the analysis.

LEMMA 2. *An inversion in the 2-sorted and 3-sorted list occurs when, and only when, the key causing the inversion is of type OEO or EOE. Hence I_n is equal to N_4, the number of keys of type $\{OEO, EOE\}$.*

The proof of this is given in Smythe and Wellner (2002).

We have now reduced the search for I_n to a simple urn model. The eight parity triples may be thought of as the states of a stochastic process. Initially the urn contains $n/3$ balls of each type (X, Y, Z). At each stage a ball is drawn at random from the urn and its type recorded, and the process continues until all the balls are drawn. At stage k, the process is in state, say, OEO, if the first k balls drawn include an odd number of X's, an even number of Y's, and an odd number of Z's. Thus N_4, which simply counts the number of times the process is in state OEO or EOE, has a direct combinatorial interpretation, and one might hope for a simple proof of its asymptotic normality. The difficulty with this urn model is that the probabilities of drawing the different types of ball at a given stage depend not just on the previous state encountered, but on the actual numbers of each type previously drawn, so the evolution of the process is complicated (and non-Markovian).

Smythe and Wellner (2002) circumvent this difficulty by a kind of Poissonization argument to define a Markov chain on a state space formed by refining the original state space to keep track of the type of ball drawn at each stage. Identifying conditional distributions resulting from this process with the distribution of N_4, the problem reduces to proving that the conditional distribution of the limit equals the limit of the conditional distribution. This seemingly obvious "fact" is not always true for non-independent summands, and the proof requires a local limit theorem for Markov chains due to Kolmogorov (1962). We finally arrive at the limiting distribution of the 1-stage of $(3, 2, 1)$-Shell Sort:

THEOREM 3. $(I_n - n/4)/\sqrt{n} \longrightarrow \lim^D \mathcal{N}(0, 3/32)$.

It would appear that the general argument outlined here would apply also to the final stage of $(h, 2, 1)$-Shell Sort for any odd h, and we would expect to get a normal limit. However, even for $h = 5$, the combinatorial difficulties are formidable, and computing the limiting variance is itself a challenge.

3. Quick Sort

Quick Sort (Hoare, 1961) is a very efficient divide-and-conquer algorithm for large data sets. Roughly, it works in the following way:

Given a list of n keys, an element is selected (by one of several means) to be called the *pivot*. Its position is located in the sorted list by comparisons with every other element in the list. As these comparisons are made, the remaining $n - 1$ elements are separated into two groups. Those greater than the pivot are moved to the right of the pivot's final position, and those less are moved to the left. The pivot is moved to its correct position between the two groups. The algorithm is then applied recursively to the two groups to the left and right of the pivot until groups of size 1 are obtained.

Quick Sort makes, on average, $2n\ln n$ comparisons to sort n keys, but its worst case behavior is $O(n^2)$. This suggests that the distribution of the number of comparisons will have a long right tail. It turns out that the limiting distribution does indeed have a long right tail, and is far from normal. There are two distinct approaches to finding a limiting distribution for Quick Sort, one using martingale theory (Regnier 1989) and one using a fixed-point argument (Rösler, 1991). The martingale method is elegant but gives little useful information about the limit distribution. The fixed-point method has proven to be useful in problems of finding a (or several) particular order statistics (Mahmoud et al (1995), Grübel and Rösler, (1996)).

The data movement to accomplish sorting, involving comparison of the pivot to the other keys, can be done in several ways; Mahmoud (2000) presents an implementation taking $n-1$ comparisons. Let C_n again denote the total number of comparisons used by Quick Sort to sort a data array of size n. The partitioning to either side of the pivot moves the pivot to a random position P_n, uniformly distributed on the integers $1, 2, \ldots, n$. The key recurrence reflecting the divide-and-conquer strategy of Quick Sort is

$$C_n \stackrel{D}{=} C_{P_n-1} + \hat{C}_{n-P_n} + n - 1, \qquad (6)$$

where $C_0 = C_1 = 0$, $\hat{C}_j \stackrel{D}{=} C_j$, and the collections $\{C_j\}$ and $\{\hat{C}_j\}$ are independent. Here the first (second) summand indicates that Quick Sort will be applied recursively to the elements to the left (right) of the pivot, and $n-1$ accounts for the partitioning stage. This relation quickly yields the expected value of C_n (Hoare, 1962):

$$E(C_n) = 2(n+1)H_n - 4n,$$

where $H_n := \sum_{j=1}^{n}(1/j)$. This gives the asymptotic relation $E(C_n) \approx 2n\ln n$. The recurrence may also be exploited to yield the variance of C_n (Knuth (1973)):

$$Var(C_n) \approx (7 - 2\pi^2/3)n^2.$$

These results suggest that the appropriate normalization for a limit theorem would be

$$C_n^* = \frac{C_n - 2n\ln n}{n}.$$

Regnier's (1989) result proves more than convergence in distribution:

THEOREM 4 $C_n^* \longrightarrow \lim^{a.s.} C$, *where C is a square-integrable random variable.*

Regnier shows that $[C_n - E(C_n)]/(n+1)$ is an L^2-bounded martingale, which therefore converges a.s. to a square-integrable random variable; Hoare's representation of $E(C_n)$ then completes the proof.

Unfortunately, this result is not easily exploited to gain information about the limiting random variable. Rösler's (1991) approach is better suited to this task. He proved that the limiting random variable C is the fixed point of a contraction mapping, using the Banach fixed-point theorem. We present a brief description of his approach.

The basic recurrence relation (6) may be manipulated to give

$$C_n^* \ =\lim^D \ \frac{P_n-1}{n}C^*_{P_n-1} \ + \ \frac{n-P_n}{n}\tilde{C}^*_{n-P_n} \ + \ G_n(P_n),$$

where

$$G_n(x)=\frac{1}{n}\big[2(x-1)\ln(x-1)+2(n-x)\ln(n-x)-2n\ln n\big] \ + \ \frac{n-1}{n},$$

and for each j, $\tilde{C}_j^* =\lim^D C_j^*$, and the collections $\{C_j^*\}$ and $\{\tilde{C}_j^*\}$ are independent. Using the fact that $(P_n-1)/n \longrightarrow \lim^D U$, where U is Uniform$(0,1)$, and overlooking some technical issues involving non-independence of the summands,, one might expect that if C_n^* converges to C, this limit should satisfy

$$C \ =\lim^D \ UC \ + \ (1-U)\tilde{C} \ + \ G(U),$$

where U is independent of C, $\tilde{C}$ is an independent copy of C, and

$$G(u) \ := \ 2u\ln u \ + \ 2(1-u)\ln(1-u)+1. \tag{7}$$

Rösler proves this convergence by a contraction argument on the space of distributions, using the Wasserstein distance of order 2 (cf., for example, Barbour, Holst and Janson (1992), who note that convergence in this metric implies the usual weak convergence of probability distributions, plus convergence of the first two moments.)

THEOREM 5. *C_n^* converges a.s. to a random variable C satisfying the distributional functional equation*

$$C \ =\lim^D \ UC \ + \ (1-U)\tilde{C} \ + \ G(U),$$

where U is Uniform$(0,1)$, U, C, and $\tilde{C}$ are independent, $\tilde{C} =\lim^D C$, and G satisfies the equation (7).

A substantial amount of work has been devoted to discerning properties of the limiting distribution C. Although no explicit form is known for it, a number of its properties have been established. In the paper establishing his proof, Rösler (1991) showed that the moment generating function existed, and Hennequin (1991) found all the cumulants of the distribution. McDiarmid and Hayward (1996) provided exponential tail bounds, confirming the right skewness; Cramer (1996) showed that the log normal provides a reasonable approximation of the distribution. Tan and Hajicostas (1995) proved that C has a density with respect to Lebesgue measure on the entire real line.

If one's interest is in finding a particular, or several, order statistics, a variant of Quick Sort, known as Find, was also developed by Hoare(1961). Probabilistic analyses of this algorithm for finding a given percentile or a random order statistic were given by Grübel and Rösler (1996) and Mahmoud, Modarres and Smythe (1995). Both of these use the recursive nature of the procedure to set up functional equations satisfied by the limit distribution, but the former problem is considerably more technical. Both papers derive some properties of the limiting distributions.

References

Barbour, A., Holst, L. and Janson, S. (1992). *Poisson Approximation.* Oxford University Press, Oxford, UK.

Cramer, M. (1996).*A Note Concerning the Limit Distribution of Quicksort Algorithm.* RAIRO: Theoretical Informatics and Its Applications , 195-207.

Grübel, R., and Rösler, U. (1996).*Asymptotic Distribution Theory for Hoare's Selection Algorithm.* Advances in Applied Probability , 252-269.

Hennequin, P. (1991). *Analyse en Moyenne d'Algorithmes, Tri Rapide, et Arbres de Recherche.* Ph.D. Dissertation, L'Ecole Polytechnique Palaiseau.

Hoare, C. (1961).*Partition (Algorithm 63), quicksort (Algorithm 64), and find (Algorithm 65). Communications of the ACM* , 3321-322.

Hoare, C. (1962). *Quicksort.* Computer Journal , 10-15.

Janson,S. and Knuth, D. (1997). *Shellsort with three Increments.* Random Structures & Algorithms , 125-142.

Johnson, B., and Killeen, T. (1983). *An Explicit Formula for the C.D.F. of the L_1 Norm of the Brownian bridge.* Annals of Probability , 807-808.

[9] Knuth, D.(1973). *The Art of Computer Programming, Vol. 1: Fundamental Algorithms*, 2nd ed. Addison-Wesley, Reading, MA.

[10] Kolmogorov, A. N. (1962). A local limit theorem for Markov chains. In *Select. Transl. Math. Statist. and Probability* **2**, 109-129. American Mathematical Society, Providence, R.I. (Translation of a Russian article *Izv.Akad. Nauk SSSR Ser. Mat.* **13**(1949), 281-300.)

[11] Lent, J., and Mahmoud, H. (1996). On tree-growing search strategies. *Annals of Applied Probability* **6**, 20-22.

[12] Louchard, G. (1986). Brownian motion and algorithmic complexity. *BIT* **26**, 17-34.

[13] Mahmoud, H. (2000). *Sorting: A Distribution Theory.* John Wiley & Sons, New York.

[14] Mahmoud, H., Modarres, R., and Smythe, R. (1995). Analysis of quickselect: An algorithm for order statistics. *RAIRO: Theoretical Informatics and Its Applications* **29**, 255-276.

[15] McDiarmid, C., and Hayward, R. (1996). Large deviation inequality for Quicksort. *Journal of Algorithms* **21**, 476-507.

[16] Régnier, M. (1989). A limiting distribution for quicksort. *RAIRO: Theoretical Informatics and Its Applications* **23**, 335-343.

[17] Rösler, U. (1991). A limit theorem for quicksort. *RAIRO: Theoretical Informatics and Its Applications* **25**, 85-100.

[18] Sedgewick, R. (1996). Analysis of Shellsort and related algorithms. *Algorithms - ESA '96: Fourth Annual European Symposium on Algorithms, Barcelona, Spain, Sept. 25-27, 1996*, J. Diaz and M. Serna, eds. Springer, Berlin, New York, 1-11.

[19] Shell, D. (1959). A high-speed sorting procedure. *Communications of the ACM* **2**, 30-32.

[20] Shepp, L. (1982). On the integral of the absolute value of the pinned Wiener process. *Annals of Probability* **10**, 234-239.

[21] Smythe, R., and Wellner, J. (2001) Stochastic analysis of shell sort. *Algorithmica* **31**, 442-457.

[22] Smythe, R., and Wellner, J. (2002) Asymptotic analysis of $(3, 2, 1)$-Shell Sort. *Random Structures & Algorithms*, to appear.

[23] Tan, K., and Hadjicostas, P. (1995). Some properties of a limiting distribution in Quicksort. *Statistics and Probability Letters* **25**, 87-94.

[24] Wellner, J. and Smythe, R. (2002). Computing the covariance of two Brownian area integrals. *Statistica Neerlandica* **56**, 101-109.

Information Functions for Reliability Analysis*

Soofi, E. S. Ebrahimi, N.

P27003

University of Wisconsin-Milwaukee, USA.
Division of Statistics, Northern Illinois University, USA.

Abstract. This expository paper presents the information theoretic framework for reliability analysis in which measuring information, inference, and information diagnostics are done in a unified manner. The whole range of available information theoretic reliability analysis methods are presented. An overview of the basic information functions, their properties, and many examples are discussed. It is our hope that this paper sets a tone for the direction of future work in the development of models for reliability using information theoretic approach.

1 Introduction

The concept of information provides a unifying theme for seemingly diverse problems and the information theory provides a unifying framework for principal activities in modeling and data analysis (Soofi 1994, 2000). In Ebrahimi and Soofi (1998), we summarized three information theoretic lines of research that have been evolved in reliability analysis. The first line of research is developing information functions that are specifically suitable for reliability analysis. The second area of research is developing various entropy-based diagnostics and tests of distributional hypothesis which are useful for reliability model building. The third information theoretic line of research, which has wide applications in reliability, is developing measures that quantify the amount of information about the immediate future contained in the past. In this paper, we update these developments in the context of reliability analysis and elaborate on these and other information theoretic developments.

More specifically, in this paper we follow a general theme that comparison of probability distributions is pivotal to statistical methodologies. In many cases the comparison is explicit. For example, hypothesis testing states the problem in terms of a comparison between the null and alternative models. However, in some cases like the maximum likelihood estimation the comparison between distributions is shown to be not so explicit (Akaike 1973, Soofi 1992). The foundation of information theoretic approaches is based on the discrimination information between probability distributions. In reliability analysis, the comparison often requires identifying the subset of the space that provides most suitable discrimination information functions for the problem at hand. We discuss various information functions in terms of discrimination information, outline their properties, and present many examples that have applications in reliability analysis.

The organization of this paper is as follows. Section 2, pertains to the Kullback-Leibler information function. Section 3 focuses on the entropy. Section 4 discusses mutual information. Section 5 focuses on the residual life distribution. Finally, Section 6 outlines some statistical applications.

* Under review for publication in *Mathematical Reliability: An Expository Perspective*, T. A. Mazzuchi, N. D. Singpurwalla, and R. Soyer (eds.), Kluwer.

2 Discrimination Information Function

Consider the problem of discriminating between two probability models F and G for a random prospect X that ranges over the space $\mathcal{S}$. Given an observation $X = x$, Bayes' theorem relates the likelihood ratio to the prior and posterior odds in favor of F as follows:

$$\log \frac{f(x)}{g(x)} = \log \frac{P(F|x)}{P(G|x)} - \log \frac{P(F)}{P(G)}, \tag{1}$$

where f and g are probability density (mass) functions, and $P(\cdot)$ and $P(\cdot|x)$ denote the prior and posterior probabilities of the model. As the difference between the posterior and prior log-odds, the logarithm of the likelihood ratio $\log[f(x)/g(x)]$ quantifies the information in $X = x$ in favor of F against G (Kullback 1959).

Suppose that x is not given, but it is known that the observation is in a set $x \in E \subseteq \mathcal{S}$ where E is not a set of measure zero. Then by taking expectation in (1),we obtain the mean information per observation $x \in E$ from F for the discrimination in favor of F against G

$$K(f : g; E) \equiv \frac{1}{P_f(E)} \int_E \left(\log \frac{f(x)}{g(x)} \right) dF(x), \tag{2}$$

where $P_f(E) = \int_E dF(x) > 0$. When there is no specific information on the whereabouts of x, other than $x \in \mathcal{S}$, the mean observation per x from F for the discrimination information between F and G is

$$K(f : g) \equiv \int \left(\log \frac{f(x)}{g(x)} \right) dF(x), \tag{3}$$

given that F is absolutely continuous with respect to G.

The discrimination information function (3) introduced by Kullback and Leibler (1951) is the fundamental information measure for comparing two distributions. It is also referred to as cross-entropy and relative entropy and generalizes two information functions, entropy and mutual information, developed by Shannon (1948). It is also interpreted as the information in F with respect to G (Savage 1954).

Suppose that $E \subseteq \mathcal{S}$ is a set of interest in a problem and the distributions of interest are the conditional (truncated) distributions $f(x|E) = f(x)/P_f(E)$ and $g(x|E) = g(x)/P_g(E)$. Then the mean information per observation x from F for the discrimination between the two conditional distributions of X, given $x \in E \subseteq \mathcal{S}$, is

$$K[f(x|E) : g(x|E)] = \frac{1}{P_f(E)} \int_E \left(\log \frac{f(x|E)}{g(x|E)} \right) dF(x) \tag{4}$$

$$= K(f : g; E) - \log \frac{P_f(E)}{P_g(E)}, \tag{5}$$

where $K(f : g; E)$ is defined in (2). That is, the discrimination information between the two conditional distributions is equal to the mean information for discrimination in favor of F against G, given E, minus the logarithm of the likelihood ratio of E under the two distributions F and G. The discrimination information function (4) is particularly useful for reliability analysis (see Section 5).

2.1 Properties of the Discrimination Information

The properties of $K(f:g)$ for discrete and continuous distributions are the same. Some properties of $K(f:g)$ are as follows (Kullback 1959):

(a) $K(f:g) \geq 0$. Equality holds if and only if $f(x) = g(x)$ almost everywhere. Therefore, $K(f:g)$ is a measure of discrepancy between the two distributions. $K(f:g)$ is not symmetric and thus is not a distance function (see Kullback 1987). It is a measure of directed divergence between f and g, where g is referred to as the *reference distribution*. The symmetric version of (3), $J(f:g) = K(f:g) + K(g:f)$, was used by Jeffreys (1946) as a measure of divergence between two distributions, but not as an information function.

(b) For mutually independent random variables $X_1, \cdots, X_n$,

$$K[f(x_1, \cdots, x_n) : g(x_1, \cdots, x_n)] = \sum_{i=1}^{n} K[f(x_i) : g(x_i)]. \tag{6}$$

(c) For any two random variables X and Y,

$$\begin{aligned} K[f(x,y) : g(x,y)] &= K[f(x) : g(x)] + E_x \{K[f(y|x) : g(y|x)]\} \\ &= K[f(y) : g(y)] + E_y \{K[f(x|y) : g(x|y)]\}. \end{aligned} \tag{7}$$

Thus, for example, $K[f(x,y) : g(x,y)] \geq K[f(x) : g(x)]$; the equality holds if and only if the expected discrimination information between the respective conditional distributions is zero.

(d) Let $E \subseteq \mathcal{S}$ with $P_g(E) > 0$. Then,

$$\int_E \left(\log \frac{f(x)}{g(x)} \right) dF(x) \geq P_f(E) \log \frac{P_f(E)}{P_g(E)}, \tag{8}$$

with equality if and only if $\dfrac{f(x)}{g(x)} = \dfrac{P_f(E)}{P_g(E)}$ for all $x \in E$. The left-hand-side of (8) is the mean information *in the elements of* E for discrimination between F and G. The right-hand-side is the discrimination information *in the set* E.

(e) Let $\{E_j,\ j = 1, \cdots, J\}$, be a partition of $\mathcal{S}$ with $\pi_j = P_f(E_j)$ and $p_j = P_g(E_j) > 0$. Let $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_J)'$ and $\boldsymbol{p} = (p_1, \cdots, p_J)'$. Then

$$K(f:g) \geq K(\boldsymbol{\pi} : \boldsymbol{p}) = \sum_{j=1}^{J} \pi_j \log \frac{\pi_j}{p_j}. \tag{9}$$

The inequalities (8) and (9) imply that grouping of observations generally lead to loss of discrimination information.

(f) Let $Y = T(X)$ be a transformation and let $f_Y(y)$ and $g_Y(y)$ denote the distributions induced by T on $f_X(x)$ and $g_X(x)$. Then $K(f_Y : g_Y) \leq K(f_X : g_X)$ with equality if and only if

$$\frac{f_Y(T(x))}{g_Y(T(x))} = \frac{f_X(x)}{g_X(x)}, \tag{10}$$

for almost all x. If condition (10) holds, T is a *sufficient statistic for discrimination*. When the distributions are indexed by a parameter θ , $K(f_Y : g_Y) \leq K(f_X : g_X)$ with equality if and only if $Y = T(X)$ is a sufficient statistic for θ.

(g) Given g, $K(f:g)$ is convex in f and given f, $K(f:g)$ is convex in g. Therefore, for a class of distributions $K(f:g)$ may be minimized with respect to either f or g, given the other.

(h) Let $f = f_\theta$ and $g = f_{\theta+\Delta\theta}$ belong to the same parametric family where θ and $\theta + \Delta\theta$ are two neighboring points in the parameter space Θ, then

$$K(f_\theta : f_{\theta+\Delta\theta}) \approx 2(\Delta\theta)^2 \mathcal{F}(\theta),$$

where

$$\mathcal{F}(\theta) \equiv \int \left[\frac{\partial \log f(x|\theta)}{\partial\theta}\right]^2 dF(x|\theta). \tag{11}$$

$\mathcal{F}(\theta)$ is Fisher information. Fisher used $\mathcal{F}(\theta)$ for the purpose of quantifying information in $f(x|\theta)$ about the parameter,a measure of information in the sense that it quantifies "the ease with which a parameter can be estimated" by x (Lehmann 1983, p. 120). Thus $\mathcal{F}(\theta)$ can be interpreted in terms of the expected information in x for discrimination between the neighboring points in Θ.

(i) An approximation given in Theil (1971) provides a Chi-square calibration. If $\boldsymbol{\pi}$ and $\boldsymbol{p}$ are two probability vectors with $\pi_j \approx p_j$ for all $j = 1, \cdots, J$, then

$$K(\boldsymbol{\pi} : \boldsymbol{p}) \approx \frac{1}{2}\sum_{j=1}^{J} \frac{(\pi_j - p_j)^2}{\pi_j}.$$

(j) The following lower bound for the discrimination information is given by Hoeffding and Wolfowitz (1958):

$$K(f:g) \geq -\log[1 - \frac{1}{4}V^2(f:g)], \tag{12}$$

where $V(f:g) = \int |f(x) - g(x)| dF(x)$ is the variation distance between two distributions. Other bounds in terms of $V(f:g)$ are given in Kullback (1967).

Often a normalized discrimination information index is needed. For the continuous case the index is computed by the following transformation of $K(f:g)$:

$$I(f:g) = 1 - e^{-K(f:g)}; \tag{13}$$

see, e.g, Soofi, Ebrahimi, and Habibullah (1995). A value $I(f:g) \approx 0$ indicates that the two distributions are very close and $I(f:g) \approx 1$ indicates that the two distributions are far apart. By (12), $\sqrt{I(f:g)} \geq V(f:g)$, which may be used for calibration of the discrimination information (Carota, Parmigiani, and Polson 1996).

2.2 Minimum Discrimination Information

The most important application of the convexity property of $K(f:g)$ is Kullback's Minimum Discrimination Information (MDI) principle for probability modeling and statistical inference. The MDI principle of modeling considers the moment class of distributions:

$$\Omega_{\boldsymbol{\theta}} = \{f(x|\boldsymbol{\theta}) : E_f[T_j(X)|\boldsymbol{\theta}] = \theta_j,\ j = 0, 1, \cdots, J\}, \tag{14}$$

where $T_j(X)$ are integrable functions with respect to the density, $T_0(x) = \theta_0 = 1$, and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_J)$ is a vector of moment parameters. Then, by the MDI Theorem (Kullback 1959, p.38), the discrimination information between any $f \in \Omega_{\boldsymbol{\theta}}$ and a distribution g is

$$K(f : g|\boldsymbol{\theta}) \geq \log \eta_0 + \boldsymbol{\eta'\theta} = K(f^* : g|\boldsymbol{\theta}), \tag{15}$$

with the equality in (15) if and only if

$$f(x|\boldsymbol{\theta}) = f^*(x|\boldsymbol{\theta}) = \eta_0 g(x) e^{\boldsymbol{\eta}' \boldsymbol{T}(x)}, \tag{16}$$

where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_J)$, $\eta_j = \eta_j(\boldsymbol{\theta})$ is the vector of Lagrange multipliers, $\boldsymbol{T}(x) = [T_1(x), \cdots, T_J(x)]'$, and $\eta_0 = \eta_0(\boldsymbol{\eta})$ is the normalizing constant. For recent developments in MDI approaches see Soofi and Retzer (2002).

2.3 Examples

Next we demonstrate some important properties of (3) through several examples.

Example 2.1: Bernoulli families

The discrimination information between two Bernoulli distributions, $\pi(x|\pi_j) = \pi_1^x \pi_2^{1-x}$, $x = 0, 1$, $\pi_1 + \pi_2 = 1$ and $p(x|p_j) = p_1^x p_2^{1-x}$, $x = 0, 1$, $p_1 + p_2 = 1$, is

$$K(\pi : p) = \pi_1 \log \frac{\pi_1}{p_1} + \pi_2 \log \frac{\pi_2}{p_2}. \tag{17}$$

McCulloch (1989) proposed the discrimination information between the flips of a fair coin and a biased coin, $K(.5 : p)$ as a calibration of $K(f : g)$. Such a calibration may be interpreted in light of (9).

(a) **Binomial distributions**
The discrimination between two binomial distributions $f(x|n, \pi)$ and $f(x|n, p)$ is

$$\begin{aligned} K(f_1 : f_2|n, \pi, p) &= n\pi_1 \log \frac{\pi_1}{p_1} + n\pi_2 \log \frac{\pi_2}{p_2} \\ &= nK(\pi : p), \end{aligned}$$

where $K(\pi : p)$ is the discrimination information between two Bernoulli distributions (17). The second equality demonstrates (6) and shows that the discrimination information between two joint distributions of the sample of independent and identical Bernoulli variables $X_1, \cdots, X_n$. We note that the discrimination information for the sample of n trials is the number of trials times the discrimination information per trial. Here, the joint discrimination is the same as that for the n-fold Bernoulli convolution.

(b) **Negative binomial distributions**
The discrimination between two negative binomial distributions $f_1(x|r,\pi)$ and $f_2(x|r,p)$ is

$$\begin{aligned} K(f_1 : f_2|r,\pi,p) &= r\log\frac{\pi_1}{p_1} + \frac{r\pi_2}{\pi_1}\log\frac{\pi_2}{p_2} \\ &= rK(\mathcal{G}_1 : \mathcal{G}_2|\pi,p) \\ &= \frac{r}{\pi}K(\pi : p), \end{aligned}$$

where $K(\mathcal{G}_1 : \mathcal{G}_2|\pi,p)$ is discrimination information between two geometric distributions and $K(\pi : p)$ is the discrimination information between two Bernoulli distributions shown in (17). . The second equality shows that the discrimination information for the negative binomial sample is the number of successes times the discrimination information for the number of trials needed for each success. The last equality shows that the discrimination information for the negative binomial sample is the expected number of trials under f_1, times the discrimination information per trial. Thus the discrimination information for the two types of sampling are the same if and only if $r/\pi_1 = n$.

(c) **Binomial with random number of trials**
Consider independent and identical Bernoulli variables $X_1, \cdots, X_N$ when N is a Poisson random variable with distributions $h_j(n|\lambda_j),\ j = 1,2$ corresponding to $\pi(x)$ and $p(x)$. Then, by (7), the joint discrimination information for the N trials is

$$\begin{aligned} K[f_1(x,n) : f_2(x,n)] &= K(h_1 : h_2|\lambda_j) + E(N|\lambda_1)K(\pi : p) \\ &= \lambda_1(\phi - \log\phi - 1) + \lambda_1 K(\pi : p), \end{aligned} \tag{18}$$

where $\phi = \lambda_2/\lambda_1$.

Example 2.2: Gamma family

Consider the gamma family with density

$$f(x|\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, \quad x > 0, \quad \lambda > 0, \alpha > 0.$$

The discrimination information between any two gamma distributions, $f_j(x|\alpha_j,\lambda_j),\ j = 1,2$, is given by

$$\begin{aligned} K(f_1 : f_2|\alpha_j,\phi) &= \left[\log\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)} + (\alpha_1 - \alpha_2)\psi(\alpha_1)\right] + \alpha_1(\phi - \log\phi - 1) \\ &= K(f_1 : f_2|\alpha_j) + \alpha_1 K(f_1 : f_2|\phi), \end{aligned}$$

where $\phi = \lambda_2/\lambda_1$ and $\psi(\cdot)$ is the digamma function. We note that the first term is the discrimination information between the shapes of the two distributions and the second term is the discrimination information between the scales of the two distributions.

Next we compute the discrimination information for some special cases of interest.

(a) **Common shape Gamma**
The discrimination information between two gamma distributions that have a common shape, $f_j(x|\alpha, \lambda_j),\ j = 1, 2$ is

$$K(f_1 : f_2|\alpha, \phi) = \alpha(\phi - \log \phi - 1) \tag{19}$$
$$= \alpha K(\mathcal{E}_1 : \mathcal{E}_2|\phi), \tag{20}$$

where $K(\mathcal{E}_1 : \mathcal{E}_2|\phi)$ is the discrimination information between two exponential distributions, $f_j(x|\lambda_j),\ j = 1, 2$. For the ease of interpretation let $\alpha = n$, an integer.

(i) Then, $f_j(x|n, \lambda_j),\ j = 1, 2$ is the density of the distribution of the waiting time to the nth event is a Poisson process with rate λ_j. Let's compare $K(f_1 : f_2|n, \phi)$ with the first term in (18), which is the discrimination information between the respective Possion distributions. The first expression (19) for $K(f_1 : f_2|n, \phi)$ is the Poisson discrimination information (18) in which the expected number of events λ_1 is replaced with the actual number of events n.

(ii) This case may also be interpreted analogously to the binomial sampling discussed above. When $\alpha = n$ is an integer, the second expression (20) is the discrimination information between the two joint distributions of the sample of independent and identical exponential variables $X_1, \cdots, X_n$. Thus, by (6), the discrimination information for the sample of n observations is the number of observations times the discrimination information per observation.

(b) **Common scale Gamma**
The discrimination information between the shapes of two gamma distributions that have a common scale, $f_j(x|\alpha_j, \lambda),\ j = 1, 2$, is

$$K(f_1 : f_2|\alpha_j) = \log \frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)} + (\alpha_1 - \alpha_2)\psi(\alpha_1).$$

Some specific cases of interest for the shape discrimination are as follows.

(i) The discrimination information between the gamma and exponential shapes is

$$K(f_1 : f_2|\alpha) = -\log \Gamma(\alpha) + (\alpha - 1)\psi(\alpha).$$

(ii) Let $\alpha_1 = \alpha_2 + 1$. Then,

$$K(f_1 : f_2|\alpha_2) = \psi(\alpha_2) - \log \alpha_2 + \frac{1}{\alpha_2}.$$

Noting that $\psi(z) - \log z \to 0$ as $z \to \infty$, for large values of α_2 there is little discrimination information, as expected. For example, the discrimination information between two chi-square distributions is

$$K(\chi^2_{n+2} : \chi^2_n) = \psi\left(\frac{n}{2}\right) - \log\left(\frac{n}{2}\right) + \frac{2}{n},$$

which vanishes as $n \to \infty$.

Example 2.3: Weibull family

Consider the Weibull family with density

$$f(x|\alpha,\lambda) = \lambda\alpha(\lambda x)^{\alpha-1}e^{-(\lambda x)^{\alpha}}, \quad x>0, \quad \lambda>0, \alpha>0.$$

The discrimination information between any two Weibull distributions, $f_j(x|\alpha_j,\lambda_j)$, $j=1,2$, let $\phi=\lambda_2/\lambda_1$, is given by:

$$K(f_1:f_2|\alpha_j,\phi) = \log\frac{\alpha_1}{\alpha_2} + \phi^{\alpha_2}\Gamma\left(1+\frac{\alpha_2}{\alpha_1}\right) - \alpha_2\log\phi + \left(1-\frac{\alpha_2}{\alpha_1}\right)\psi(1) - 1$$

A special case of interest is the discrimination information between the general Weibull and exponential shapes, given by

$$K(f_1:f_2|\alpha) = \log\alpha + \Gamma\left(1+\frac{1}{\alpha}\right) + \left(1-\frac{1}{\alpha}\right)\psi(1) - 1.$$

Note that the functional relationship between the exponential and Weibull variables constrains the relationship between the scale parameters such that $\phi=1$ when $\lambda_1=\lambda_2=1$.

Example 2.4: Systems of components

Consider a system that consists of n components $X_1,\ X_2,\ldots,X_n$. Here $X_1,\ X_2,\ldots,X_n$ are assumed to be independent with common density f, survival function $\bar{F}$ and the cumulative distribution function F. Consider another system with n components $Y_1,\ldots,Y_n$. Here $Y_1,\ldots,Y_n$ are assumed to be independent with common density g, survival function $\bar{G}$ and the cumulative distribution G.

(a) **Series components**
It is clear that the lifetime of the first system is $Z_1=\min(X_1,\ X_2,\cdots,X_n)$ with the density function

$$f_{Z_1}(z) = n\ f(z)\ (\bar{F}(z))^{n-1}.$$

Also, the lifetime of the second system is $Z_2=\min(Y_1,\cdots,Y_n)$ with the density function

$$f_{Z_2}(z) = n\ g(z)\ (\bar{G}(z))^{n-1}.$$

Now,

$$K(f_{Z_1}:f_{Z_2}) = E_{f_{Z_1}}\left(\log\frac{f(Z)}{g(Z)}\right) + (n-1)E_{f_{Z_1}}\left(\log\frac{\bar{F}(Z)}{\bar{G}(Z)}\right). \qquad (21)$$

(b) **Parallel components**
In this case, $Z_1=\max(X_1,\cdots,X_n)$, $Z_2=\max(Y_1,\cdots,Y_n)$, $f_{Z_1}(z)=nf(z)\ F^{n-1}(z)$ and $f_{Z_2}(z)=ng(z)\ G^{n-1}(z)$, and

$$K(f_{Z_1}:f_{Z_2}) = E_{f_{Z_1}}\left(\log\frac{f(Z)}{g(Z)}\right) + (n-1)E_{f_{Z_1}}\left(\log\frac{F(Z)}{G(Z)}\right).$$

Note that in this case it is clear from equation (21) that if components lifetimes are closed (not distinguishable), then the systems lifetimes will be also closed.

3 Entropy

Shannon's entropy (Shannon 1948) defined by

$$H(f) \equiv H[f(x)] = -\int \log f(x) dF(x) \tag{1}$$

is the entropy of f.

Entropy may be computed using the hazard (failure) rate function. For a non-negative random variable X

$$H(f) = 1 - \int (\log \lambda_F(x)) dF(x), \tag{2}$$

where $\lambda_F(x) = \dfrac{f(x)}{\bar{F}(x)}$ is the hazard rate, $\bar{F}(x) = 1 - F(x)$ denotes the survival (reliability) function (Teitler, Rajagopal, and Ngai 1986).

The entropy is a discrimination information function in the following sense:

$$K(f : U) = H(U) - H(f), \tag{3}$$

where U denotes the uniform distribution. (For simplicity, a finite support is assumed in (3). The infinite support case may be written in terms of limits.)

By the entropy-information relation (3), the negative entropy $-H(f)$ measures lack of uniformity (concentration of probabilities) under f. With a less concentrated distribution, it is more difficult to predict an outcome. Thus, $-H(f)$ is a measure of informativeness of f about the prediction of its outcomes (Zellner 1971). This is in accordance with the fact that $-H(f) = E_f[\log f(X)]$ is the average log-height of the density. Accordingly, a distribution f_1 is more informative than f_2 if and only if

$$\Delta H(f_2, f_1) \equiv H(f_2) - H(f_1) = K(f_1 : U) - K(f_2 : U) \geq 0.$$

For a set of interest $E \subseteq \mathcal{S}$, we may define the entropy of the set $H(f : E)$ and the entropy of the conditional (truncated) distribution $H[f(x|E)]$ similarly to (2) and (4).

For two random variables X_1 and X_2, with the joint density $f(x_1, x_2)$ the joint entropy, $H(X_1, X_2)$, and the partial entropy of X_2 given a particular value $X_1 = x_1$ denoted by $H(X_2|x_1)$, are obtained by replacing $f(x)$ in (1) by $f(x_1, x_2)$ and the conditional density $f(x_2|x_1)$, respectively. The expected entropy of the conditional density

$$H(X_2|X_1) = \int H(X_2|x_1) dF_1(x_1) \tag{4}$$

is generally referred to as the *conditional entropy* of X_2 given X_1.

3.1 Properties of Entropy

(a) In the discrete case with a finite support, entropy is non-negative, but the entropy of a continuous random variable X takes values in $[-\infty, +\infty]$.

(b) In the discrete case, entropy is invariant under one-to-one transformations of X, but the entropy of a continuous random variable X is not invariant under one-to-one transformations of X. If $Y = T(X)$ is a transformation function, then

$$H(Y) = H(X) - E\left[\log\left|\frac{d}{dX}T^{-1}(Y)\right|\right]. \tag{5}$$

(c) For mutually independent random variables $X_1, \cdots, X_n$,

$$H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i). \tag{6}$$

(d) The entropy of an n-dimensional random vector $(X_1, \cdots, X_n)'$ is decomposable as

$$H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}), \tag{7}$$

where $H(X_i|X_1, \cdots, X_{i-1})$ is the conditional entropy of X_i given the other variables. Hence, in general decomposition of a joint entropy is order-dependent.

(e) The entropy $H(f)$ is concave in f.

3.2 Maximum Entropy

One of the principal activities in science is assessing distribution functions and random prospects based on partial information. The *Maximum Entropy (ME)* principle of scientific inference (Jaynes 1957) serves this purpose.

The ME distribution is the one that maximizes (1) with respect to f subject to the set of constraints in (14), which reflect the partial knowledge about f. The ME model in (14) is given by (15) with $g(x) = 1$,

$$f^*(x|\boldsymbol{\theta}) = \eta_0 e^{\boldsymbol{\eta}'\boldsymbol{T}(x)}, \tag{8}$$

Therefore, for any $f \in \Omega_{\boldsymbol{\theta}}$,

$$H[f(x|\boldsymbol{\theta})] \leq -\log\eta_0 - \boldsymbol{\eta}'\boldsymbol{\theta} = H[f^*(x|\boldsymbol{\theta})], \tag{9}$$

where the equality is attained by the ME distribution (8). The MDI inequality (15) generalizes the entropy inequality (9) by a sign reversal. Note that when g is uniform in $K(f : g)$, by the entropy-information relation (3), the MDI and ME procedures are equivalent. The MDI is referred to as the *Minimum Cross-Entropy* principle. For the rationale and axiomatic justifications of the ME and MDI principles see Jaynes (1968), Shore and Johnson (1980) and Csiszar (1991).

An interesting and very important ME result is obtained when F is a distribution with support of the entire real line and $\Omega_\theta = \left\{f(x|\theta) : E|X|^k \leq \theta\right\}$. Then for $\theta < \infty, k > 0$, we have the following entropy-moment inequality:

$$H(X) \leq \frac{1}{k}\log\frac{2^k e\Gamma^k(1/k)\theta^k}{k^{k-1}}. \tag{10}$$

The equality in (10) is attained by the density

$$f^*(x) = \eta_0 e^{-\eta|x|^k}, \quad \eta > 0. \tag{11}$$

Note that $k = 1$ gives double-exponential (Laplace) distribution and $k = 2$ give normal (Gaussian) distribution.

Table 1. Examples of Entropy and Variance Orderings of Distributions.

Family & Density	Variance	Entropy	Ordering
Exponential $f(x) = \frac{e^{-\lambda x}}{\lambda}$	$\frac{1}{\lambda^2}$	$-\log\lambda + 1$	$\downarrow \lambda$
Gamma $f(x) = \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}$	α	$\log\Gamma(\alpha) + (1-\alpha)\psi(\alpha) + \alpha$	$\uparrow \alpha$
Inverse Gamma $f(x) = \frac{e^{-x^{-1}}x^{-\alpha-1}}{\Gamma(\alpha)}$	For $\alpha > 2$, $\frac{1}{(\alpha-1)^2(\alpha-2)}$	$\log\Gamma(\alpha) - (1+\alpha)\psi(\alpha) + \alpha$	$\downarrow \alpha > 2$
Generalized-normal $f(x) = \frac{x^{\alpha-1}e^{-x^2}}{2\Gamma(\alpha/2)}$	$\frac{\alpha - 2\Gamma^2(\alpha/2+1/2)}{2\Gamma^2(\alpha/2)}$	$\log\frac{\Gamma(\alpha/2)}{2} + \frac{\alpha+(1-\alpha)\psi(\alpha/2)}{2}$	$\uparrow \alpha$
Pareto $f(x) = \alpha x^{-\alpha-1},\ x > 1$	For $\alpha > 2$, $\frac{1}{\alpha(\alpha-1)^2(\alpha-2)}$	$-\log\alpha + \frac{1}{\alpha} + 1$	$\downarrow \alpha > 2$
Weibull $f(x) = \alpha x^{\alpha-1}e^{-x^\alpha}$	$\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)$	$-\log\alpha + \frac{(\alpha-1)\gamma}{\alpha} + 1$ $\gamma = 0.5772\cdots$	$\downarrow \alpha > \gamma$
Beta $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$	$\log[B(\alpha,\beta)]$ $-(\alpha-1)[\psi(\alpha) - \psi(\alpha+\beta)]$ $-(\beta-1)[\psi(\beta) - \psi(\alpha+\beta)]$	$\uparrow \alpha, (\alpha,\beta) \in R_\alpha$ $\downarrow \alpha, (\alpha,\beta) \in S_\alpha$ $\uparrow \beta, (\alpha,\beta) \in R_\beta$ $\downarrow \beta, (\alpha,\beta) \in S_\beta$

Definitions of the regions R_α, S_α, R_β, and S_β:

$$R_\alpha = \left\{ (\alpha,\beta) : \alpha < \frac{\sqrt{(\beta+1)(9\beta+1)} - (\beta+1)}{4} \right\} \quad S_\alpha = \left\{ (\alpha,\beta) : \alpha > 1 - \frac{(\beta-1)\psi_\alpha(\alpha+\beta)}{\psi_\alpha(\alpha) - \psi_\alpha(\alpha+\beta)} \right\}$$

$$R_\beta = \left\{ (\alpha,\beta) : \beta < \frac{\sqrt{(\alpha+1)(9\alpha+1)} - (\alpha+1)}{4} \right\} \quad S_\beta = \left\{ (\alpha,\beta) : \beta > 1 - \frac{(\alpha-1)\psi_\beta(\alpha+\beta)}{\psi_\beta(\beta) - \psi_\beta(\alpha+\beta)} \right\}$$

3.3 Entropy and Variance Orderings

Ebrahimi, Maasoumi and Soofi (1999a,b) have discussed ordering of distributions on the basis of variance and entropy and given conditions where these two ordering are equivalent. Unlike variance which measures concentration only around the

mean, entropy measures diffuseness of the density irrespective of the location(s) of concentration. Table 1 gives examples of variance and entropy orderings of some well known distributions extracted from Ebrahimi, Maasoumi and Soofi (1999a). The arrows in Table 1 indicate the entropy and variance in the family decrease ($\downarrow$) or increase ($\uparrow$) with the parameter. As shown in Table 1, for most distributions variance and entropy order the distributions over the parameter space similarly. However, we note that this is not always true. For example, variance and entropy order the Weibell family differently when $\alpha < \gamma$, where γ is the Euler number. For gamma, inverse-gamma, Pareto, and Weibull, the scale parameter $\lambda = 1$. By (5), $H(\lambda X) = H(X) - \log \lambda$, hence both entropy and variance are decreasing in λ. The case of beta family is more complicated. The two measures order the beta family similarly only over certain regions in the parameter space, $R_\alpha, R_\beta, S_\alpha$, and S_β defined in the table.

Furthermore, for a distribution the variance may not be defined when the entropy is finite, e.g., inverse gamma and Pareto distributions shown in the table. However, as may be seen from the entropy-moment inequality (10), with $k = 2$ a finite entropy is implied by a finite variance.

4 Mutual Information

The information provided by a particular value $X = x$ about Y is given by the amount of uncertainty difference

$$\Delta H[f(y), f(y|x)] = H(Y) - H(Y|x).$$

The uncertainty difference is positive (negative) when conditional density $f(y|x)$ is farther (closer) to uniformity than the marginal density $f_2(y)$.

The mutual information (Shannon 1948) is defined by the expected entropy difference,

$$\begin{aligned} M(X,Y) &\equiv E_x[H(Y) - H(Y|x)] && (1)\\ &= H(Y) - H(Y|X) && (2)\\ &= H(Y) + H(X) - H(X,Y) && (3)\\ &= K[f(x,y) : f_1(x)f_2(y)], && (4) \end{aligned}$$

where $H(Y|X)$ is the conditional entropy defined in (4).

The mutual information is Shannon's measure of the expected information about the input signal Y transmitted through a noisy channel which transmits the output signal X. As is apparent in (4), $M(X,Y) = M(Y,X) \geq 0$ and $M(X,Y) = 0$ if and only if two variables are independent. Thus, $M(X,Y)$ measures the extent of functional dependency between X and Y

Normalized mutual information indices between two variables are obtained as follows. For the continuous case, (13) is used,

$$I(X,Y) \equiv I[f(x,y) : f_1(x)f_2(y)] = 1 - e^{-2M(X,Y)}. \quad (5)$$

For the discrete case,

$$I(X,Y) \equiv I[f(x,y) : f_1(x)f_2(y)] = \frac{M(X,Y)}{H_M}, \tag{6}$$

where H_M denotes a base uncertainty, e.g., $H_M = \max\{H(X), H(Y)$, or $H_M = H(Y)$ if Y is the variable about which the uncertainty reduction is of interest. Note that $I(X,Y) = 0$ if and only if the two variables are independent, and $I(X,Y) = 1$ if and only if the two variables are functionally dependent (Joe 1989).

Mutual information between a random variable Y and an n-dimensional random vector $\boldsymbol{X}$ is

$$\begin{aligned} M(Y,\boldsymbol{X}) &= H(Y) - H[Y|(X_1,\cdots,X_n)] \\ &= \sum_{i=1}^{n} M(Y,X_i|X_1,\cdots,X_{i-1}), \end{aligned}$$

where $M(Y,X_i|X_1,\cdots,X_{i-1})$ is the partial mutual information. The latter is interpreted as a measure of the remaining dependency between Y and X_i after controlling for the dependency between Y and $(X_1,\cdots,X_{i-1})$.

4.1 Expected Information About Parameter

When X is data and Y is a random parameter θ with a prior distribution over Θ, then the mutual information referred to as Lindley's measure of information in the data about about the parameter (Lindley 1956), given by

$$M(\Theta,X) = H(\Theta) - H(\Theta|X) \tag{7}$$

$$= E_x\{K[f(\theta|x) : f(\theta)]\}. \tag{8}$$

Lindley's measure has been successfully applied in developing information loss (gain) diagnostics for experimental design, censoring, and many other statistical problems, see Ebrahimi and Soofi (1998) and references therein.

Bernardo (1979) provided an expected utility interpretation of $M(\Theta,X)$ based on (8), which is now prevalent in Bayesian literature. The *reference* priors are obtained by maximizing $M(\Theta,X)$ with respect to $f(\theta)$. In general, maximization of $M(\Theta,X)$ does not give an explicit solution. Lindley (1961) showed that ignorance between two neighboring values θ and $\Delta\theta$ in the parameter space implies that $M(\Theta,X) \approx 2(\Delta\theta)^2\mathcal{F}(\theta)$, where $\mathcal{F}(\theta)$ is Fisher information.

In (7) the expectation is taken with respect to the marginal distribution $f(x)$. Zellner (1971) defined an information function as the difference between the prior entropy and the entropy of the sampling distribution (likelihood), averaged with respect to the prior distribution $E_\theta H[f(x|\theta)]$. Zellner's measure of information in the data about the parameter is

$$\begin{aligned} Z(\Theta) &\equiv H(\Theta) - H(X|\Theta) \\ &= E_\theta\{K[f(x|\theta) : f(\theta)]\}. \end{aligned} \tag{9}$$

This function is not a mutual information. The second expression gives an interpretation of $Z(\Theta)$ in terms of the discrimination information function between the

likelihood and the prior. Zellner (1997) gave a new interpretation of $Z(\Theta)$ in terms of "the *total* information provided by an experiment over and above the prior" defined as $I(Exp) \equiv H(\Theta) - H(\Theta, X) = -E_\theta[H(X|\theta)]$, where $H(\Theta, X)$ is the entropy of the joint distribution. The Maximal Data Information Prior (MDIP) maximizes $Z(\Theta)$. The MDIPs are in the form of $p(\theta) \propto \exp\{-H[f(y|\theta)]\}$.

Example 4.1: Pedagogical Example

Consider three bivariate distributions shown in Table 2. For each case, the conditional distributions $f(y|x),\ x = 0, 1, 2$ and their entropies are also shown in the right panel of the Table. The information measures for the three distributions are shown in the lower panel of the table. In all three cases, $H(Y) = .67$.

Table 2. Information measures for three bivariate distributions.

Joint and marginal distributions | Entropies of conditional distributions

a) $f(x, y)$ general bivariate distribution

	x			
y	0	1	2	$f_2(y)$
0	0	.25	.15	.40
1	.25	.25	.10	.60
$f_1(x)$	.25	.50	.25	

	x		
y	0	1	2
0	0	.50	.60
1	1	.50	.40
$H(Y\|x)$	0	.69	.67

b) $f(x, y)$ when X and Y are independent

	x			
y	0	1	2	$f_2(y)$
0	.10	.20	.10	.40
1	.15	.30	.15	.60
$f_1(x)$	.25	.50	.25	

	x		
y	0	1	2
0	.40	.40	.40
1	.60	.60	.60
$H(Y\|x)$	.67	.67	.67

c) $f(x, y)$ when X and Y are related functionally

	x			
y	0	1	2	$f_2(y)$
0	.20	0	.20	.40
1	0	.60	0	.60
$f_1(x)$	.20	.60	.20	

	x		
y	0	1	2
0	1	0	1
1	0	1	0
$H(Y\|x)$	0	0	0

Information measure	(a) General	(b) Independent	(c) Related Functionally
Marginal Entropy $H(Y)$	.67	.67	.67
Conditional Entropy $H(Y\|X)$	.52	.67	0
Mutual Information $M(X, Y)$	.15	0	.67
Information Index $I(X, Y)$	23%	0%	100%

In the general case (a), we note that $H(Y) > H(Y|x = 0),\ H(Y) < H(Y|x = 1)$,

and $H(Y) = H(Y|x = 2)$. That is, $x = 0$ reduces uncertainty (is informative) about outcome of Y, $x = 1$ increases uncertainty about outcome of Y, and $x = 2$ leaves the uncertainty unchanged. However, $H(Y) > H(Y|X)$, indicating that, on average, knowledge of X reduces uncertainty about outcomes of Y. The mutual information $M(X, Y) = .15$ quantifies this average uncertainty reduction. The fraction of uncertainty reduction, computed using the normalized index (6) with $H_M = H(Y)$, is $I(X, Y) = 23\%$.

In case (b), the bivariate distribution has the independent structure. The entropies of the conditional distributions (shown in the right panel of the Table) are all equal to the marginal entropy $H(Y) = .67$. Thus, the amount of average uncertainty is $H(Y|X) = .67$, there is no uncertainty reduction, and $I(X, Y) = 0\%$.

In case (c), the variables are related functionally, $P(Y = 2X - X^2) = 1$. The entropies of the conditional distributions (shown in the right panel of the Table) are all zero, hence no uncertainty about the outcomes remains when an $X = x$ is available for the prediction of Y. Therefore, the mutual information is equal to the marginal entropy $M(X, Y) = H(Y) = .67$ and the normalized is $I(X, Y) = 100\%$. However, we note that the correlation coefficient $\rho(X, Y) = 0$ due to the fact that the functional relationship between Y and X is nonlinear.

Example 4.2: To survive or to fail?

Abel and Singpurwalla (1994) considered the lifetime of an item X with the exponential distribution with mean $E(X|\theta) = \theta$ and failure rate λ and posed the following questions: Which of the two outcomes, *survival* or *failure*, in a small interval $(t_0, t_0 + \Delta t_0)$ is more informative about θ and λ? They provided an answer using a gamma prior for λ with density $p(\lambda|\alpha, \beta) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$. The implied prior for the mean $\theta = 1/\lambda$ is inverse gamma. The posterior distributions for $\lambda \in \Lambda$ and $\theta \in \Theta$ are gamma and inverted gamma with a shape parameter $\alpha(\alpha + 1)$ and a scale parameter $(t_0 + \beta)$.

The information provided by the data (survival or failure) about λ and θ are quantified by the differences between entropies of respective prior and posterior distributions:

$$\mathcal{I}_\lambda(survival) = H(\Lambda) - H(\Lambda|survival) = \log \frac{t_0 + \beta}{\beta} \tag{10}$$

$$\mathcal{I}_\lambda(failure) = H(\Lambda) - H(\Lambda|failure) = \log \frac{t_0 + \beta}{\beta} + (H_\alpha - H_{\alpha+1}) \tag{11}$$

$$\mathcal{I}_\theta(survival) = H(\Theta) - H(\Theta|survival) = \log \frac{\beta}{t_0 + \beta} \tag{12}$$

$$\mathcal{I}_\theta(failure) = H(\Theta) - H(\Theta|failure) = \log \frac{\beta}{t_0 + \beta} + \left(H_\alpha - H_{\alpha+1} + \frac{2}{\alpha}\right), \tag{13}$$

where H_α is the entropy of the gamma distribution with shape parameter α and scale equal one, shown in Table 1.

As shown in Table 1, H_α is increasing in α, so $H_\alpha - H_{\alpha+1} < 0$, which implies that $\mathcal{I}_\lambda(survival) > \mathcal{I}_\lambda(failure)$. Abel (1991) has shown that for $\alpha \geq 1$, the quantity $H_\alpha - H_{\alpha+1} + 2/\alpha > 0$, which implies that $\mathcal{I}_\theta(failure) > \mathcal{I}_\theta(survival)$. Thus, a

survival is more informative about the failure rate λ and a failure is more informative about the mean θ.

Also note that $\mathcal{I}_\lambda(survival) = -\mathcal{I}_\theta(survival) < 0$. Furthermore, $\mathcal{I}_\lambda(survival)$ is increasing in t_0 and $\mathcal{I}_\theta(survival)$ is decreasing in t_0. For any $t_0 > 0$, $\mathcal{I}_\lambda(survival) > 0$ and $\mathcal{I}_\theta(survival) < 0$. That is, a survival is always informative about the failure rate, but always increases uncertainty about the mean. The case is not so clear for $\mathcal{I}_\lambda(failure)$ and $\mathcal{I}_\theta(failure)$.

The expected information about the parameters are given by the mutual information:

$$M(\Lambda, Y) = \mathcal{I}_\lambda(survival)P(survival) + \mathcal{I}_\lambda(failure)P(failure) > 0$$
$$M(\Theta, Y) = \mathcal{I}_\theta(survival)P(survival) + \mathcal{I}_\theta(failure)P(failure) > 0,$$

where Y is a binary random variable that indicates survival and failure in interval $(t_0, t_0 + \Delta t_0)$.

Finally, let $\hat{\lambda}$ and $\hat{\theta}$ be the Maximum Likelihood Estimate (MLE) of λ and θ. Then, Fisher information (11) gives the following results: $\mathcal{F}_{\hat{\lambda}}(survival) = t_0^2 > 0$ and is increasing in t_0; $\mathcal{F}_{\hat{\theta}}(survival) = t_0^{-2} > 0$ and is decreasing in t_0; but $\mathcal{F}_{\hat{\lambda}}(failure) = \mathcal{F}_{\hat{\theta}}(failure) = 0$. Thus, Fisher information (11) does not provide a meaningful answer to the question of interest in this problem when a failure occurs in $(t_0, t_0 + \Delta t_0)$.

Example 4.3: Multivariate normal

Consider $(Y, X_1, \cdots, X_p)$ with $(p+1)$-variate normal distribution.

(a) Let $\boldsymbol{X} = (X_1, \cdots, X_p)'$. Then $\boldsymbol{X}$ has p-variate normal distribution with covariance $\Sigma_X = [\sigma_{ij}]$.

(i) The entropy of conditional distribution $f(x_i|x_j)$ is

$$H(X_i|X_j = x_j) = \frac{1}{2}\log(2\pi e) + \frac{1}{2}\log[(1-\rho_{ij}^2)\sigma_{ii}],$$

where ρ_{ij} is the correlation coefficient. Thus, $H(X_i|X_j = x_j)$ is a function of the correlation and variance and is independent of the value x_j.

(ii) The mutual information between any pair (X_i, X_j), $i \neq j$ is just the entropy difference,

$$\begin{aligned} M(X_i, X_j) &= H(X_i) - H(X_i|X_j = x_j) \\ &= -\frac{1}{2}\log(1-\rho_{ij}^2) \geq 0. \end{aligned}$$

The mutual information index is $I(X_i, X_j) = \rho_{ij}^2$. This reflects the fact that for the multivariate normal variables, stochastic dependency and linear dependency among $X_1, \cdots, X_p$ are equivalent.

(iii) The mutual information between the vector $\boldsymbol{X}$ and its components is given by:

$$M[\boldsymbol{X}, (X_1, \cdots, X_p)] = \frac{1}{2}\sum_{j=1}^{p} \log \sigma_{jj} - \frac{1}{2}\log |\Sigma_X|$$
$$= \frac{1}{2}\sum_{j=1}^{p} \log \sigma_{jj} - \frac{1}{2}\sum_{\ell=1}^{p} \log \lambda_\ell,$$

where $|\Sigma_X|$ denotes the determinant, and λ_ℓ is the ℓth eigenvalue of Σ_X. The normalized index of dependency (5) is

$$I[\boldsymbol{X}, (X_1, \cdots, X_p)] = 1 - \frac{\lambda_1 \cdots \lambda_p}{\sigma_{11} \cdots \sigma_{pp}}.$$

Linear dependency is indicated by some $\lambda_\ell = 0$, which leads to $I[\boldsymbol{X}, (X_1, \cdots, X_p)] = 1$. On the other extreme, $\lambda_j = \sigma_{jj},\ j = 1, \cdots, p$ if and only if $X_1, \cdots, X_p$ are mutually uncorrelated (independent) for which $I[\boldsymbol{X}, (X_1, \cdots, X_p)] = 0$.

(b) It can be shown that for any set of given $x_1, \cdots, x_p$, the entropy of conditional distribution $H[Y|(x_1, \cdots, x_p)]$ is a function of the variances and covariances, and is functionally independent of $(x_1, \cdots, x_p)$, and

$$\Delta\{H(Y), H[Y|(x_1, \cdots, x_p)]\} = -\frac{1}{2}\log[1 - \rho^2(Y; X_1 \cdots X_p)] \geq 0, \qquad (14)$$

where $\rho^2(Y; X_1 \cdots X_p)$ is the square of the multiple correlation between Y and $X_1, \cdots, X_p$. Therefore by (14), any set of multivariate normal data $(x_1, \cdots, x_p)$ is informative about Y.

(c) In this case, the mutual information is given by the entropy difference:

$$M(Y, X_1, \cdots, X_p) = E_{\boldsymbol{X}}[\Delta\{H(Y), H[Y|(x_1, \cdots, x_p)]\}]$$
$$= -\frac{1}{2}\log\left[1 - \rho^2(Y; X_1 \cdots X_p)\right]$$
$$= -\frac{1}{2}\sum_{i=1}^{p} \log\left[1 - \rho^2(Y; X_1 \cdots X_i)\right],$$

where $\rho^2(Y; X_1 \cdots X_p)$ is the square of the multiple correlation between Y and $X_1, \cdots, X_p$ and $\rho^2(Y; X_1 \cdots X_i)$ is the square of the partial correlation between Y and X_i, given $X_1, \cdots, X_{i-1}$. The normalized index of dependency (5) is given by the square of the multiple correlation $I[Y, (X_1, \cdots, X_p)] = \rho^2(Y; X_1 \cdots X_p)$.

5 Information in Residual Lifetime

Frequently, in reliability one has information about the current age of the system under consideration. In such cases, the age must be taken into account when measuring information. Ebrahimi and Kirmani (1996b,c) considered the situations when age t must be taken into account.

If we think of a random variable X as the lifetime of a system then X is a non-negative random variable. In this case, the set of interest is the *residual life*, $E_t = \{x : x > t\}$

5.1 Residual Discrimination Information

Ebrahimi and Kirmani (1996b,c) proposed using the discrimination information function between two residual life distributions for the system $F_t(x) = P(X - t \leq x|X > t)$ and $G_t(x) = P(X - t \leq x|X > t)$ implied by two lifetime distributions $F(x)$ and $G(x)$. The discrimination information between the two residual life distributions is given by:

$$K(f : g; t) \equiv K(f_t : g_t) = \int_t^\infty f_t(x) \log \frac{f_t(x)}{g_t(x)} dx \tag{1}$$

$$= K(f : g; E_t) - \log \frac{\bar{F}(t)}{\bar{G}(t)}, \tag{2}$$

where $f_t(x) = f(x)/\bar{G}(t)$ and $g_t(x)/\bar{F}(t)$ denote the conditional densities, $\bar{F}(t) = P_f(E_t) = 1 - F(t)$ and $\bar{G}(t) = P_g(E_t) = 1 - G(t)$ are the survival functions, and $K(f : g; E_t)$ is defined in (2). It is clear that for $t_0 = inf\{x : F(x) = 1\}$, $K(f : g; t_0) = K(f, g)$.

By (2), the discrimination information between two residual distributions is equal to the mean information for discrimination in favor of F against G, given E_t, minus the logarithm of the likelihood ratio of of the survival of the system beyond t under the two lifetime distributions F and G. By (4), for each t, $t \geq 0$, $K(f : g; t)$ possesses all the properties of the discrimination information function (3). If we consider t as an index ranging over E_t, then $K(f : g; t)$ provides a dynamic discrimination information function indexed by t for measuring the discrepancy between the residual life distributions $F_t(x)$ and $G_t(x)$. It can be shown that $K(f : g; t)$ is free of t if and only if the hazard functions are proportional, i.e., $\bar{G}(t) = \bar{F}^\beta(t), \beta > 0$. For more details see Ebrahimi and Kirmani (1996b,c).

The following example demonstrates computation and usefulness of $K(f, g; t)$.

Example 5.1: Systems of components

Consider again the systems of n components discussed in Example 2.4.

(a) **Series components**
For $Z_1 = \min(X_1, X_2, \cdots, X_n)$ and $Z_2 = \min(Y_1, \cdots, Y_n)$, the dynamic discrimination information is given by

$$K(f_{Z_1} : f_{Z_2}; t) = E_{f_{Z_1|Z_1>t}} \left(\log \frac{f(Z)}{g(Z)} \right) + (n-1) E_{f_{Z_1|Z_1>t}} \left(\log \frac{\bar{F}(Z)}{\bar{G}(Z)} \right) + n \log \frac{\bar{G}(t)}{\bar{F}(t)}.$$

(b) **Parallel components**
For $Z_1 = \max(X_1, X_2, \cdots, X_n)$ and $Z_2 = \max(Y_1, \cdots, Y_n)$, the dynamic discrimination information is given by

$$K(f_{Z_1} : f_{Z_2}; t) = E_{f_{Z_1|Z_1>t}} \left(\log \frac{f(Z)}{g(Z)} \right) + (n-1) E_{f_{Z_1|Z_1>t}} \left(\log \frac{F(Z)}{G(Z)} \right) + \log \frac{1 - [G(t)]^n}{1 - [F(t)]^n}.$$

Note that in both cases $K(f_{Z_1} : f_{Z_2}; 0) = K(f_{Z_1} : f_{Z_2})$ computed in Example 2.4.

5.2 Residual Entropy

The entropy of residual life distribution is defined similarly as

$$H(X;t) \equiv H(f;t) = -\int_t^\infty \frac{f(x)}{\bar{F}(t)} \log \frac{f(x)}{\bar{F}(t)} dx. \tag{3}$$

As in (2), $H(f;t)$ may be computed using the hazard (failure) rate function,

$$H(f;t) = 1 - \frac{1}{\bar{F}(t)} \int_t^\infty f(x) \log \lambda_F(x) dx.$$

It is clear that for $t_0 = inf\{x : F(x) = 1\}$, $H(f;t_0) = H(f)$.

Like the entropy (1), the residual entropy (3) is a discrimination information function as in (3). As before, let $U(x)$ denote the uniform distribution with support $\mathcal{S} = \{x : a < x < b\}$. Then conditional distribution of X, given $x > t$ is also uniform, i.e., $U_t(x)$ is uniform over (t, b) and

$$K(f : U;t) = H(U;t) - H(f;t). \tag{4}$$

Therefore $H(f;t)$ measures the uncertainty (or lack of predictability) of the remaining life-time of a system of age t. It can be shown that the dynamic entropy $H(f;t)$, like the failure rate and mean residual life, uniquely determines the distribution function F. On the basis of the measure $H(f;t)$, we can define some non-parametric classes of life distributions that are closely related to other classes of life distributions, such as increasing failure rate (IFR) and decreasing failure rate (DFR).

A survival function $\bar{F}$ ($\bar{F}(0) = 1$) is said to have decreasing (increasing) uncertainty or residual life (DURL (IURL)) if $H(f;t)$ is decreasing (increasing) in t. One can easily show that for the exponential distribution $H(f;t)$ remains constant. That is, uncertainty about lifetime does not change as the system ages. In fact, the exponential distribution is both DURL and IURL. For further properties and implications of $H(f;t)$, DURL and IURL classes see Ebrahimi (1996) and Ebrahimi and Kirmani (1996a).

In reliability, there are many situations in which the hazard rate function $\lambda_F(t)$ must satisfy certain constraints. In fact, we argue that state of no knowledge about physical characteristic of a system at all is hardly, if ever realistic and we would typically at least have some idea concerning physical behavior of a system. Ebrahimi, Hamadani, and Soofi (1996) studied developing lifetime distribution through maximizing the entropy $H(f)$ subject to monotonocity constraints on failure rate $\lambda_F(t)$. However, to produce a model for the data generating distribution function f under these constraints the direct use of $H(f,t)$ is more appropriate. Because, given $X > t$ we are interested in modeling distribution of X_t, the remaining lifetime of a system of age $t \geq 0$. When partial information is available about $\lambda_F(t)$ we can develop a model for X_t by maximizing $H(f;t)$ instead of $H(f)$ in the ME problem; see Ebrahimi (2000).

5.3 Residual Mutual Information

We may define a dynamic version of mutual information by

$$M(Y,X;t_1,t_2) = K[f(x,y) : f_1(x)f_2(y); t_1, t_2] \tag{5}$$

$$= K[f(x,y) : f_1(x)f_2(y); E_{t_1,t_2}] - \log \frac{\bar{F}(t_1,t_2)}{\bar{F}_1(t_1)\bar{F}_2(t_2)} \tag{6}$$

where $E_{t_1,t_2} = \{(x,y) : x > t_1, y > t_2\}$ the set of interest in this problem and

$$K[f(x,y) : f_1(x)f_2(y); E_{t_1,t_2}] = \frac{1}{\bar{F}(t_1,t_2)} \int_{t_1}^{\infty} \int_{t_2}^{\infty} \log \left[\frac{f(x,y)}{f_1(x)f_2(y)} \right] f(x,y)\ dxdy.$$

It is clear that for $t_{j0} = inf\{x : F_j(x) = 1\},\ j = 1,2,\ M(Y,X;t_{10},t_{20}) = M(Y,X)$. The dynamic mutual information measures the extent of functional dependence of remaining lifetimes of two systems that are already survived t_1 and t_2 respectively. It is clear that if $M(Y,X;t_1,t_2) = 0$, then the residual lifetimes are independent.

Note that the dynamic mutual information $M(Y,X;t_1,t_2)$ is defined in terms of (4). In terms of the marginal and joint entropies $M(Y,X;t_1,t_2)$ is given by

$$M(Y,X;t_1,t_2) = \frac{\bar{F}_1(t_1)\bar{F}_2(t_2)}{\bar{F}(t_1,t_2)}[H(Y;t_1) + H(X;t_2)] - H(X,Y;t_1,t_2). \tag{7}$$

This is analogous to (3). It is possible to express $M(Y,X;t_1,t_2)$ in terms of the entropy difference and conditional entropy analogously to (1) and (2).

The following example demonstrates computation and usefulness of $M(X,Y;t_1,t_2)$.

Example 5.2:

Consider a system that consists of two components. Suppose that the first component has lifetime X, the second component has lifetime Y, and X and Y has the Basu and Block's joint density, see Block and Basu (1974),

$$f(x,y) = \begin{cases} \dfrac{\lambda_1\lambda(\lambda_2+\lambda_{12})}{\lambda_1+\lambda_2}\ \exp(-\lambda_1 x - (\lambda_2+\lambda_{12})y), & x < y \\ \dfrac{\lambda_2\lambda(\lambda_1+\lambda_{12})}{\lambda_1+\lambda_2}\ \exp(-(\lambda_1+\lambda_{12})x - \lambda_2 y), & x > y \end{cases}$$

where $\lambda = \lambda_1 + \lambda_2 + \lambda_{12}$. The variables are independent when $\lambda_{12} = 0$.

The joint survival function is

$$\bar{F}(t_1,t_2) = \frac{\lambda}{\lambda_1+\lambda_2} \exp\{-\lambda_1 t_1 - \lambda_2 t_2 - \lambda_{12} \max(t_1,t_2)\} - \frac{\lambda_{12}}{\lambda_1+\lambda_2} \exp\{-\lambda \max(t_1,t_2)\}, \tag{8}$$

The marginal densities and survival functions are, for $k = 1,2$,

$$f_k(x) = \frac{\lambda(\lambda_k+\lambda_{12})}{\lambda_1+\lambda_2}\ \exp\{-(\lambda_k+\lambda_{12})x\} - \frac{\lambda_{12}\lambda}{\lambda_1+\lambda_2} \exp\{-\lambda x\},$$

$$\bar{F}_k(x) = \frac{\lambda}{\lambda_1+\lambda_2}\ \exp\{-(\lambda_k+\lambda_{12})x\} - \frac{\lambda_{12}}{\lambda_1+\lambda_2} \exp\{-\lambda x\}.$$

For $t_1 = t_2 = t$ the bivariate survival function (8) simplifies to $\bar{F}(t,t) = \exp\{-\lambda t\}$. In order to simplify the computation, we let $t_1 = t_2 = t$. From equation (2.7),

$$M(X,Y;t,t) = \log \frac{\bar{F}_1(t)\bar{F}_2(t)}{\bar{F}(t,t)} + \frac{1}{\bar{F}(t,t)} \int_t^\infty \int_t^\infty f(x,y) \log f(x,y) dx\ dy$$

$$- \frac{1}{\bar{F}(t,t)} \sum_{k=1}^{2} \bar{F}_{3-k}(t) \int_t^\infty f_k(x) \log f_k(x) dx.$$

One can show that:

$$\log \frac{\bar{F}_1(t)\bar{F}_2(t)}{\bar{F}(t,t)} = -2\log(\lambda_1+\lambda_2) + \log[\lambda^2 \exp\{-\lambda_{12}t\} - \lambda\lambda_{12} \exp\{-(\lambda_1+\lambda_{12})t\}$$
$$-\lambda\lambda_{12} \exp\{-(\lambda_2+\lambda_{12})t\} + \lambda_{12}^2 \exp\{-\lambda t\}].$$

$$\int_t^\infty \int_t^\infty (log\ f(x,y)) f(x,y)\ dx\ dy = \left(\frac{\lambda_1 \log a_1 + \lambda_2 \log a_2}{\lambda_1+\lambda_2} - \lambda t - 2 \right) e^{-\lambda t},$$

where $a_k = \dfrac{\lambda_k \lambda(\lambda_{3-k}+\lambda_{12})}{\lambda_1+\lambda_2},\ k = 1, 2.$

$$\int_t^\infty f_k(x) \log f_k(x) dx = \bar{F}_k(t) \log \frac{\lambda(\lambda_k+\lambda_{12})}{\lambda_1+\lambda_2} - \left[\frac{(\lambda_k+\lambda_{12})\lambda t + \lambda}{\lambda_1+\lambda_2} \right] e^{-(\lambda_k+\lambda_{12})t}$$
$$+ \frac{\lambda_{12}(\lambda_k+\lambda_{12})(\lambda t+1)}{\lambda(\lambda_1+\lambda_2)} e^{-\lambda t}$$
$$- \sum_{k=1}^{\infty} \left(\frac{\lambda_{12}}{\lambda_k+\lambda_{12}} \right)^k \frac{\lambda(\lambda_k+\lambda_{12})}{k(\lambda_1+\lambda_2)(\lambda_k+k\lambda_{3-k}+\lambda_{12})} e^{-(\lambda_k+k\lambda_{3-k}+\lambda_{12})t}$$
$$+ \sum_{k=1}^{\infty} \left(\frac{\lambda_{12}}{\lambda_k+\lambda_{12}} \right)^k \frac{\lambda\lambda_{12}}{k(\lambda_1+\lambda_2)(k\lambda_{3-k}+\lambda)}\ e^{-((k\lambda_{3-k}+\lambda))t}.$$

Here if we put $t = 0$, then the result coincides with the one given by Ahsanullah and Habibullah (1996).

Suppose that $\lambda_1 = \lambda_2 = \lambda_{12} = 1$, then

$$M(X,Y;t,t) = -2\log 2 + \log\left[9e^{-t} - 6e^{-2t} + e^{-3t}\right]$$
$$+(\log 3 - 2 - 3t)e^{-t} \left[\{3t - 6(e^{-t}-1)(\log 3 - 1)\} e^{-2t} - \left(\frac{1}{2}\log 3 - \frac{1}{3} - t \right) e^{-3t} \right.$$
$$\left. -3\sum_{k=1}^{\infty} \frac{1}{2^k k(k+2)} e^{-(k+2)t} + 3\sum_{k=1}^{\infty} \frac{1}{2^{k+1} k(k+3)} e^{-(k+3)t} \right].$$

We note that as $t \to \infty$, $M(X, Y; t, t) \to 0$, indicating that as the components age the dependency between their remaining lives are weakened.

6 Information Statistics

In this section, we outline some statistical applications of the information functions discussed above. This line of research still provides numerous rich theoretical and applied problems for the future.

6.1 Covariate Information Index

Covariate information indices are measures that quantify the impact of a set of variables $\boldsymbol{X} = (X_1, \cdots, X_p)$ on the distribution of another variable Y. The most well-known measure of covariate information is the R^2 of regression. The R^2 is meaningfully interpretable for the Gaussian case. This is shown to be the case in various information theoretic formulations. We have seen in Example 4.3 that when the variables are jointly normal, the conditional entropy $H(Y|x_1, \cdots, x_p)$ is functionally independent of $x_1, \cdots, x_p$. In this case, the mutual information is just the entropy difference, $M(Y, \boldsymbol{X}) = H(Y) - H(Y|x_1, \cdots, x_p)$, and the sample counterpart of the normalized mutual information $I(Y, \boldsymbol{X})$ is the R^2 of regression.

A formulation that is particularly useful for reliability analysis is the generalization of R^2 in the context of the exponential family regression. Consider the regression problem $E(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ is the random vector of responses, $\boldsymbol{X} = [x_{ij}]$ is an $n \times p$ matrix of given covariate values x_{ij}, and $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression parameters. Suppose that the distribution of $\boldsymbol{Y}$ has a density in the exponential family

$$f_{\boldsymbol{\eta}}(\boldsymbol{y}) = h(\boldsymbol{y}) \exp\{\boldsymbol{\eta}'\boldsymbol{y} - \Psi(\boldsymbol{\eta})\},$$

where $\boldsymbol{\eta} \in \mathcal{H}$ is the vector of natural or canonical parameters, $\Psi(\boldsymbol{\eta})$ is the normalizing function, and $h(\boldsymbol{y})$ is a parameter-free function.

When the covariance matrix is positive definite, the relation between $\boldsymbol{\eta} \in \mathcal{H}$ and $E(\boldsymbol{Y}) = \boldsymbol{\mu} \in \mathcal{M}$ is one-to-one. For any member of the exponential family, the discrimination information has the sequential additive properties for nested linear subspaces of the natural parameter space $\mathcal{H}$ and the expectation parameter space $\mathcal{M}$ (Kullback 1971, Simon 1973). The covariate information index for exponential family regression is derived based on the additivity of information in the natural parameter space.

Let $\boldsymbol{\eta}_r = \boldsymbol{X}_r\boldsymbol{\beta}_r$, where $\boldsymbol{X}_r$ is an $n \times r$ full-rank matrix and $\boldsymbol{\eta}_s = \boldsymbol{X}_s\boldsymbol{\beta}_s$, where $\boldsymbol{X}_s$ is an $n \times s$ full-rank submatrix of $\boldsymbol{X}_s$, $r \leq s \leq n$. Then

$$K(f_{\boldsymbol{\eta}^*} : f_{\boldsymbol{\eta}_s^*}) = K(f_{\boldsymbol{\eta}^*} : f_{\boldsymbol{\eta}_r^*}) + K(f_{\boldsymbol{\eta}_r^*} : f_{\boldsymbol{\eta}_s^*}), \tag{1}$$

where $\boldsymbol{\eta}^*$, $\boldsymbol{\eta}_r^*$, and $\boldsymbol{\eta}_s^*$ are the MDI estimates (Simon (1973) used the MLE). Hastie (1997) formulated the exponential family regression estimation in terms of (1) with

$$\boldsymbol{\eta}^*|\boldsymbol{y} = \boldsymbol{\eta}^*(\boldsymbol{y}), \quad \boldsymbol{\eta}_r^*|\boldsymbol{y} = \boldsymbol{\eta}_r^*(\bar{y}), \text{ and } (\boldsymbol{\eta}_r^*|\boldsymbol{y} = \hat{\mu}) = \boldsymbol{\eta}_r^*(\boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Covariate information for exponential family regression is

$$I_{\bar{y}}(\boldsymbol{X}) = \frac{K(f_{\hat{\mu}} : f_{\bar{y}})}{K(f_y : f_{\bar{y}})} = 1 - \frac{K(f_y : f_{\hat{\mu}})}{K(f_y : f_{\bar{y}})}$$

For the normal regression problem, $I_{\bar{y}}(\boldsymbol{X}) = R^2$. More relevant distributions for reliability applications are exponential and gamma. Cameron and Windmeijer (1997) tabulated the covariate index $I_{\bar{y}}(\boldsymbol{X})$ for several distributions in the exponential family. For example, the index for the Gamma family is

$$I_{\bar{y}}(\boldsymbol{X}) = 1 - \frac{\sum \log(y_i/\hat{\mu}) + (y_i/\hat{\mu}) - 1}{\sum \log(y_i/\bar{y})}.$$

6.2 Distributional Fit Diagnostics

The discrimination information function and entropy have been instrumental in the development of indices of fit of parametric models to the data. Given data $x_1, \cdots, x_n$ from a distribution F, it is important to assess whether the unknown $F(x)$ can be satisfactorily approximated by a parametric model $F^*(x|\boldsymbol{\theta})$. The loss of approximating $F(x)$ by a parametric model $F^*(x|\boldsymbol{\theta})$ is measured by $K(f : f^*|\boldsymbol{\theta})$. In order to assess the loss of approximation the unknown data-generating distribution $F(x)$ by a model $F^*(x|\boldsymbol{\theta})$, the discrimination information $K(f : f^*|\boldsymbol{\theta})$ must be estimated. In general, estimation of (3) directly is formidable.

Akaike considered approximating the unknown data-generating distribution $f(x)$ by a family of models $f^*(\boldsymbol{x}|\boldsymbol{\theta_J})$ and estimating the model parameter $\boldsymbol{\theta}_J$, including its dimension J, $J = 1, \cdots, L$. Akaike (1973) showed that "choice of the information theoretic loss function is a very natural and reasonable one to develop a unified asymptotic theory of estimation." The approximation loss is measured by the information discrepancy $K[f(\boldsymbol{x}) : f^*(\boldsymbol{x}|\tilde{\boldsymbol{\theta}})]$. The MDI or minimum relative entropy loss estimate of $\boldsymbol{\theta}$ is defined by

$$\tilde{\boldsymbol{\theta}}_{MDI} = \arg\max_{\boldsymbol{\theta}} K[f(\boldsymbol{x}) : f^*(\boldsymbol{x}|\theta)]. \tag{2}$$

The entropy loss has been used with frequentist and Bayesian risk functions in various parametric estimation problems and for model selection; see Soofi (1997) and references therein.

Akaike (1974) observed that decomposing the log-ratio in (3) gives

$$K(f : f^*|\boldsymbol{\theta}) = -E_f[\log f^*(X|\boldsymbol{\theta})] - H[f(x)], \tag{3}$$

where $H[f(x)]$ is the entropy of f. Since the entropy of the data-generating distribution is free of the parameters, the second term in (3) is ignored in the derivation of the AIC for model selection. The minimization of the information discrepancy between the unknown data-generating distribution and the model is operationalized by maximizing the average log-likelihood function in (3).

Since the AIC type measures are derived by minimizing the first term in (3) and the second term is ignored, the AIC type measures provide criteria for model comparison purposes only, and do not provide information diagnostic about the model fit. An alternative approach for estimating $K(f : f^*|\boldsymbol{\theta})$ when f is an unknown

distribution is proposed by Soofi et al. (1995). They considered the moment class (14) and showed that if $f \in \Omega_{\boldsymbol{\theta}}$ and f^* is the ME model in $\Omega_{\boldsymbol{\theta}}$, then the first term in (3) becomes the entropy of f^* and

$$K(f : f^*|\boldsymbol{\theta}) = H[f^*(x|\boldsymbol{\theta})] - H[f(x|\boldsymbol{\theta})]. \tag{4}$$

This equality defines the information distinguishability (ID) between distributions in $\Omega_{\boldsymbol{\theta}}$. The first term is the entropy of the parametric model ME (8) and the second term is the entropy of a distribution which is unknown other than the density is a member of a general moment class (14). Like the Akaike's decomposition (3), this decomposition proves to be quite useful for developing model selection criteria. Ebrahimi (2001a) has given other conditions where equivalence of two entropies implies that two distributions are equal.

ID statistics are obtained by estimating (3) via (4). The normalized ID index for the continuous case is computed as:

$$\begin{aligned} ID(f_n : f^*|\boldsymbol{\theta}_n) &= 1 - \exp[-K(f_n : f^*|\boldsymbol{\theta}_n)] \\ &= 1 - \exp\{H[f_n(x|\boldsymbol{\theta}_n)] - H[f^*(x|\boldsymbol{\theta}_n)]\}, \end{aligned} \tag{5}$$

where f_n is a nonparametric estimate with entropy $H[f_n(x|\boldsymbol{\theta}_n)]$ and moments $\boldsymbol{\theta_n} = (\theta_{1,n}, \cdots, \theta_{A,n})'$, and f^* is the ME model in $\Omega_{\boldsymbol{\theta}_n}$. An $ID(f_n : f^*|\boldsymbol{\theta}_n) = 0$ indicates the perfect fit; i.e., f^* is a perfect parameterization of f_n. A lower bound for (5) in terms of variation distance is given by (12), $ID(f_n : f^*|\boldsymbol{\theta}_n) \geq \frac{1}{4}V^2(f_n : f^*|\boldsymbol{\theta}_n)$.

Implementation of ID indices of fit includes two steps. First, a parametric model $f^*(x|\boldsymbol{\theta})$ is selected based on the maximum entropy characterization of the densities of the parametric families. Many commonly known parametric families are shown to admit ME characterization. On the other hand, for a parametric model, one may easily identify the moment class $\Omega_{\boldsymbol{\theta}}$ by writing the density in the exponential form (8). The entropy expression (9) for the well known parametric families are tabulated, see, e.g., Soofi et al. (1995). The second step for implementation of ID indices is the nonparametric estimation of $H[f_n(x|\boldsymbol{\theta}_n)]$. Various nonparametric entropy estimates for continuous distributions are developed in the literature which can be used as $H(f_n)$ in (5). However, maintaining the non-negativity of the estimate of (5) is an important issue. For this purpose, the parameters of the maximum entropy $H[f^*(x|\boldsymbol{\theta}_n)]$ in (5) *must* be estimated by the moments of the density f_n whose entropy is $H[f_n(x|\boldsymbol{\theta}_n)]$; for references and more details, see Soofi and Retzer (2002).

For example, many current results in life testing are based on the assumption that the life of a system is described by an exponential distribution. Of course, in many situations this assumption is usually suspect. When F is a distribution with support on the positive real line, then the exponentional distribution is the ME model $f^*(x|\boldsymbol{\theta})$ in the moment class,

$$\Omega_\theta = \{f(x|\theta) : E(X) \leq \theta\}.$$

We can estimate θ using the mean θ_n of a nonparametric density estimate f_n with entropy $H(f_n)$ and compute the ID statistic $ID(f_n : f^*|\boldsymbol{\theta}_n)$. Then for large (small) values of $ID(f_n : f^*|\boldsymbol{\theta}_n)$ we reject (accept) the exponential model for the data.

Developing ME fit indices is a very promising line of research. Recent developments includes Ebrahimi (1998, 2001b) who uses dynamic discrimination information function for developing tests of exponentiality and uniformity of the residual

lifetime, and Mazzuchi et al. (2000, 2002) who develop Bayesian estimation and inference about entropy and the model fit.

6.3 Statistical Process Control

Alwan, Ebrahimi and Soofi (1998) using information functions (3) and (1) proposed information theoretic process control (ITPC) as a framework which formalizes some current SPC practices and broadens the scope of the SPC so that various types of process parameters can be monitored in a unified manner. In the ITPC framework, they developed signal charting procedures for monitoring of various types of moments without a need for making distributional assumptions. The most important feature of ITPC is that various monitoring problems are handled in a unified manner based upon a criterion function (3).

The ITPC procedure for monitoring of moments consists of a three-step algorithm. In the first step, the in-control moment values $\boldsymbol{\theta_0} = (\theta_{10}, \cdots, \theta_{J0})$ are the only available information. The n-control moment values are used as inputs to the ME procedure which produces a model $f_0^*(x|\boldsymbol{\theta_0})$ for the unknown distribution the process variable X.

The second step is for estimating the distribution of the process variable X at the monitoring state. At each stage $t = 1, 2, \cdots$ the information at hand are the ME model for the in-control distribution and the data moments $\boldsymbol{m_t} = (m_{1t}, \cdots, m_{Jt})$. The MDI algorithm uses moments m_{jt} (new information) and the initial ME model $f_0^*(x|\boldsymbol{\theta_0})$ as the inputs, minimizes $K(f_t, f_0^*)$ with respect to f_t and produces a new model $f_t^*(x|\boldsymbol{m_t})$ for the distribution of X at the monitoring state.

The third step is for detecting a change in the distribution of the process variable between monitoring state and the in-control state. The process is monitored based on the MDI function $K(f_t^*, f_0^*)$. The final step is the most important feature of the ITPC algorithm for monitoring moments because it solves the traditional problem of constructing charts based on problem specific statistical criterion functions deemed suitable for the problem at hand.

Alwan et al (1998) derived various MDI functions for ITPC charts, developed examples of MDI control charts for the multivariate case and process attribute, and discuss possibility of developing control charts for detecting distributional change by application of (5).

As an example, Alwan et al (1998) examined the performance of the Information Chart for monitoring mean and variance of the process variable. For monitoring of mean and variance, the conventional SPC assumption of normality in not needed. Whence the in-control parameters $\boldsymbol{\theta} = (\mu_0, \sigma_0^2)$ are given, the model $f^*(x|\mu_0, \sigma_0^2) = N(\mu_0, \sigma_0^2)$ is found as the ME solution. At the monitoring stage, using the sample mean and variance $\boldsymbol{m_t} = (\bar{x}_t, s_t^2)$ the MDI control function for the detecting mean and/or variance shifts is

$$\begin{aligned} IMV_t = 2nK_t(f_t^* : f_0^*|\mu_0, \sigma^2, \bar{x}_t, s_t^2) &= \frac{n(\bar{x}_t - \mu_0)^2}{\sigma_0^2} + n\left[\frac{s_t^2}{\sigma_0^2} - \log\frac{s_t^2}{\sigma_0^2} - 1\right] \\ &= IM_t + IV_t. \end{aligned}$$

The first term in IM_t measures the *information discrepancy* due to the process mean and the second term IV_t measures the *information discrepancy* due to the process variation.

The IM-chart constructed by plotting IM_t is equivalent to the Shewhart mean chart. Using μ_0 for the mean in the monitoring state instead of $\bar{x}_t$ gives $IM_t = 0$ and we obtain the MDI control function for the process variance, IV_t, shown in IV_t. Note that the term s_t^2/σ_0^2 of IV_t is the chi-square statistic associated with the s^2 control chart that detects shifts in process dispersion. Thus, IMV_t embraces two control charting procedures traditionally used for mean and variance as its special cases.

6.4 Prediction Problems

The entropy-moment equality (10) is shown to play a key role in the prediction problem (Shepp, Slepian, and Wyner 1980). Let $\hat{Y}$ denote a predictor of Y. Then (10) provides a sharp lower bound for the prediction error variance in terms of the entropy of $X = Y - \hat{Y}$:

$$E(Y - \hat{Y})^2 \geq \frac{e^{2H(Y-\hat{Y})}}{2\pi e}, \tag{6}$$

with equality holding when $Y - \hat{Y}$ is Gaussian. Using (6), Pourahmadi and Soofi (2000) developed a sharp lower bound for the prediction error variance of non-Gaussian ARMA processes. Their lower bound is for the variance of any unbiased predictor.

As an example, consider the $ARMA(1,1)$ model

$$Y_t - \phi Y_{t-1} = Z_t + \theta Z_{t-1}, \quad \phi + \theta \neq 0, \quad |\phi| < 1, \quad |\theta| < 1,$$

where $\{Z_t\}$ is a sequence of i.i.d. random variables with mean zero and variance σ^2, called the *innovation process*. The model is stationary. For prediction of Y_0 on the past $Y_t,\ t = -1, -2, \cdots$, a result of Pourahmadi and Soofi (2000) gives

$$E|Y_0 - \hat{Y}_0|^2 \geq \frac{e^{2H(Z_0)}}{2\pi e} \frac{\log|\theta|}{\log|\phi|}, \tag{7}$$

with equality holding for Gaussian processes. Conceptually, the role of entropy in (7) is the role of the inverse of Fisher information which provides a lower bound for the variance of an unbiased estimator via the Cramer-Rao inequality.

When the distribution of the innovation in the ARMA process is known, $H(Z_0)$ can be estimated parametrically. However, the more realistic and interesting case is when the innovation distribution is unknown. Then a nonparametric estimate of the entropy can be used as a yardstick against which to gauge the fits of various competing parametric models assessed through their one step ahead prediction error variances. For Gaussian data details of this idea has been worked out in Mohanty and Pourahmadi (1996) and references therein. For the non-Gaussian case, a nonparametric estimate of the innovation process entropy is needed.

References

Abel, P.S. (1991). Information and the Design of Life Tests. unpublished Ph.D. dissertation, George Washington University. Department of Operation Research.

Abel, P.S., and Singpurwalla, N.D. (1994). To Survive or to Fail: That is the Question. *The American Statistician*, **48**, 18-21.

Akaike, H. (1974) "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.*, AC-19, 716-723.

Akaike, H. (1973) "Information Theory and an Extension of the Maximum Likelihood Principle", *2nd International Symposium on Information Theory*, 267-281.

Alwan, L.C., Ebrahimi, N., and Soofi, E. (1998). Information theoretic framework for process control. *European Journal of Operational Research*, **111**, 526-542.

Bernardo, J. M. (1979), "Expected Information as Expected Utility", *Annals of Statistics*, 7, 686-690.

Block, H.B. and Basu, A. P. (1974). "A Continuous Bivariate Exponential Extension", *Journal of the American Statistical Association*, 69, 1031-1037.

Cameron, A. C. and F. A. G. Windmeijer (1997), "An R-squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models", *Journal of Econometrics*, 77, 329-342.

Carota, C., G. Parmigiani, and N. G. Polson (1996) "Diagnostic Measures for Model Criticism", *Journal of the American Statistical Association*, 91, 753-762.

Csiszar, I. (1991), "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference in Linear Inverse Problems", *Ann. Statist.*, 19, 2032-66.

Ebrahimi, N. (2001a). Families of distributions characterized by entropy. *IEEE Trans. on Information Theory*, **97**, 2042-2045.

Ebrahimi, N. (2001b). Testing for uniformity of the residual lifetime based on dynamic Kullback-Leibler information. *Annals of the Inst. of Stat. Math.*, **53**, 325-337.

Ebrahimi, N. (2000). The maximum entropy method for lifetime distributions. *Sankhya* A, **62**, 236-243.

Ebrahimi, N. (1998). Testing exponentiality of the residual life, based on dynamic Kullback-Leibler information. *IEEE Trans. on Reliability*, **47**, 197-201.

Ebrahimi, N. (1996). How to measure uncertainty in the residual lifetime distributions. *Sankhya* A, **58**, 48-57.

Ebrahimi, N. and Kirmani, S. (1996a). Some results on ordering of survival functions through uncertainty. *Stat. and Probab. Letters*, **29**, 167-176.

Ebrahimi, N. and Kirmani, S. (1996b). A characterization of the proportional hazards model through a measure of discrimination between two residual life distributions. *Biometrika*, **83**, 233-235.

Ebrahimi, N. and Kirmani, S. (1996c). A measure of discrimination between two residual lifetime distributions and its applications. *Ann. Inst. Statist. Math*, **48**, 257-265.

Ebrahimi, N. and Soofi, E. (1998). Recent developments in information theory and reliability analysis. *Frontiers in Reliability*, 125-132. Edited by A.P. Basu, S.K. Basu and S. Mukhopadhyay, World Scientific, New Jersey.

Ebrahimi, N., Hamadani, G.G., and Soofi E.S. (1991) "Maximum Entropy Modeling with Partial Information on Failure Rate", School of Business Administration, University of Wisconsin-Milwaukee.

Ebrahimi, N., Maasoumi, E., and Soofi, E. S. (1999a) "Ordering Univariate Distributions by Entropy and Variance", *Journal of Econometrics*, 90, 317-336.

Ebrahimi, N., Maasoumi, E., and Soofi, E. S. (1999b) "Measuring Informativeness of Data by Entropy and Variance", in *Advances in Econometrics: Income Distribution and Methodology of Science, Essays in Honor of Camilo Dagum*, D. Slottje (ed), 61-77, New York: Physica-Verlag.

Hastie, T. (1987) " A Closer Look at the Deviance", *The American Statistician*, 41, 16-20.

Hoeffding, W. and Wolfowitz, J. (1958), "Distinguishability of Sets of Distributions", *Annals of Mathematical Statistics*, 29, 700-718.

Jaynes, E.T. (1968) "On the Rationale of Maximum-Entropy Methods," *Proceedings of IEEE*, 70, 939-952.

Jaynes, E.T. (1957). Information theory and statistical mechanics. *Physics Review*, **106**, 620-630.

Jeffreys, H. (1946) "An Invariant Form for the Prior Probability in Estimation Problems", *Proceedings of Royal Statistical Society (London), A*, 186, 453-461.

Joe, H. (1989) "Relative Entropy Measures of Multivariate Dependence", *Journal of the American Statistical Association*, 84,

Kullback, S (1987), "The Kullback-Leibler Distance", *The American Statistician*, 41, 340.

Kullback, S (1971), "Marginal Homogeneity of Multidimensional Contingency Tables", *The Annals of Mathematical Statistics*, 42, 594-606.

Kullback, S (1967), "A Lower Bound for Discrimination Information in Terms of Variation", *IEEE Transactions on Information Theory*, IT-13,

Kullback, S. (1959), *Information Theory and Statistics*, N.Y.: Wiley (reprinted in 1968 by Dover).

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *The Annals of Math. Stat.*, **22**, 79-86.

Lehmann, E. L. (1983), *Theory of Point Estimation*, N.Y.: Wiley.

Lindley, D. V. (1961), "The Use of Prior Probability Distributions in Statistical Inference and Decision", *Proceedings of the Fourth Berkeley Symposium*, 1, 436-468, Berkeley: UC Press.

Lindley, D.V. (1956). On a measure of information provided by an experiment. *The Annals of Math. Stat.*, **27**, 986-1005.

Mazzuchi, T. A., Soofi, E.S. and Soyer, R. (2002), "Bayes Estimate and Inference for Entropy and Information Index of Fit", submitted for publication.

Mazzuchi, T. A., Soofi, E.S. and Soyer, R. (2000), "Computations of Maximum Entropy Dirichlet for Modeling Lifetime Data", *Computational Statistics and Data Analysis*, 32, 361-378.

McCulloch, R. E. (1989), "Local Model Influence", *Journal of the American Statistical Association*, 84, 473-478.

Mohanty, R. and Pourahmadi, M. (1996), "Estimation of the generalized prediction error variance of a multiple time series," *Journal of the American Statistical Association*, 91, 294-299.

Pourahmadi, M. and Soofi, E. S. (2000), "Predictive Variance and Information Worth of Observations in Time Series", *Journal of Time Series Analysis*, 21, 413-434.

Savage, L. J. (1954). *The Foundations of Statistics*, New York: John Wiley.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379-423.

Shepp, L.A., Slepian, D. and Wyner, A.D. (1980), "On Prediction of Moving-Average Processes", *The Bell System Technical Journal*,

Shore, J. E., and R. W. Johnson (1980), "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," *IEEE Transac. on Info. Theory*, IT-26, 26-37.

Simon, G. (1973), "Additivity of Information in Exponential Family Probability Laws", *Journal of the American Statistical Association*, 68, 478-482.

Soofi, E. S. (2000) "Principal Information Theoretic Approaches" *Journal of the American Statistical Association*, 95, 1349-1353.

Soofi, E. S. (1997), "Information Theoretic Regression Methods", in *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems*, 12, T. B. Fomby and R. C. Hill (eds.), 25-83, Greenwich, CT: JAI Press.

Soofi, E. S. (1994) "Capturing the intangible concept of Information", *Journal of the American Statistical Association*, 89, 1243-1254.

Soofi, E. S. (1992) "A Generalizable Formulation of Conditional Logit With Diagnostics", *Journal of the American Statistical Association*, 87, 412-816.

Soofi, E. S., and Retzer, J.J. (2002) "Information Indices: Unification and Applications", *Journal of Econometrics*, 107, 17-40.

Soofi, E. S., Ebrahimi, N, and Habibullah, M. (1995), "Information Distinguishability with Application to Analysis of Failure Data", *Journal of the American Statistical Association*, 90, 657-668.

Theil, H. (1971), *Principles of Econometrics*, New York: John Wiley.

Teitler, S., Rajagopal, A.K., and Ngai, K.L. (1986), "Maximum Entropy and Reliability Distributions", *IEEE Transactions on Reliability* R-35, 391-395.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley (reprinted in 1996 by Wiley)

Zellner, A. (1997), *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*, Cheltenham UK: Edward Elgar.

Information Properties of Order Statistics and Spacings

Soofi, E. S. and Ebrahim, N.

P27003

University of Wisconsin-Milwaukee, USA.
Division of Statistics, Northern Illinois University, USA.

Abstract. Order statistics are used in analysis of various problems in many fields. However, thus far not much has been explored about the information properties of the order statistics and spacings between order statistics. We explore information properties of order statistics based on the entropy, Kullback-Leibler information, and mutual information. The probability integral transformation plays a pivotal role in developing our results. We provide bounds for the entropy of order statistics and some results that relate entropy ordering among order statistics to other well known orderings of random variables. We show that the discrimination information between order statistics and data distribution, the discrimination information among the order statistics, and the mutual information between order statistics are all distribution free and are computable using the distributions of the order statistics of the samples from the uniform distribution. In the final section, we discuss information properties of spacings for uniform and exponential samples and provide a large sample distribution free result on the entropy of spacings. The results show interesting symmetries of information orderings among order statistics which confirm intuition and provide useful insights about the information properties of order statistics.

1 Introduction

Suppose that $X_1, \cdots, X_n$ are independent and identically distributed observations from an absolutely continuous distribution F_X with density f_X. The order statistics of the sample is defined by the arrangement of $X_1, \cdots, X_n$ from the smallest to the largest, denoted as $Y_1 < \cdots < Y_n$. It is well known that the distribution $F_i(y) = P(Y_i \leq y)$ has the following density:

$$f_i(y) = \frac{\Gamma(n+1)}{\Gamma(n-i+1)\Gamma(i)}[F_X(y)]^{i-1}[1-F_X(y)]^{n-i}f_X(y), \quad i = 1, \cdots n, \qquad (1.1)$$

where for a positive integer z, $\Gamma(z) = (z-1)!$ is the gamma function.

Order statistics have been used in a wide range of problems, including in robust statistical estimation and detection of outliers, characterization of probability distributions and goodness-of-fit tests, entropy estimation, analysis of censored samples, reliability analysis, quality control, strength of materials, waiting time until a big event, selecting the best, records and allocation of prize in tournaments, inequality measurement, speech processing, image and picture processing, echo removal, image coding, filtering, spectrum estimation, acoustics, and edge enhancing; see Arnold, Balakrishnan, and Nagaraja (1992), Wong and Chen (1990), Ebrahimi, Pflughoeft, and Soofi (1994), and references therein. In spite of such a wide scope of applications, not much attention has been given to the study of information properties of order statistics. We have been able to find only three papers on this topic. Wong

and Chen (1990) showed that the difference between the average entropy of order statistics and the entropy of data distribution is a constant. They also showed that for symmetric distributions, the entropy of order statistics is symmetric about the median. Park (1995) showed some recurrence relations for the entropy of order statistics and Park (1996) provided similar results in terms of the Fisher information.

In this paper, we develop several results on the properties of the entropy of order statistics and on the Kullback-Leibler discrimination information functions that involve order statistics. The probability integral transformation of the random variable, $U = F_X(X)$, plays a pivotal role in developing our results. It is well known that the distribution of U is uniform over the unit interval. The order statistics of a sample from uniform distribution $U_1, \cdots, U_n$ are denoted by $W_1 < \cdots < W_n$ and W_i has beta distribution with density

$$g_i(w) = \frac{1}{B(i, n-i+1)} w^{i-1}(1-w)^{n-i}, \quad 0 \le w \le 1, \quad i = 1, \cdots n, \tag{1.2}$$

where $B(z_1, z_2) = \Gamma(z_1)\Gamma(z_2)/\Gamma(z_1 + z_2)$.

This paper is organized as follows. Section 2 presents some results on the entropy of order statistics. Section 3, gives some results on the discrimination information function related to order statistics. Section 4 presents an asymptotic result on the entropy of the spacings.

2 Entropy of order statistics and its Properties

The probability integral transformation provides the following useful representation of entropy of the random variable X,

$$\begin{aligned} H(X) &= -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx \\ &= -\int_0^1 \log f_X\left(F_X^{-1}(u)\right) du. \end{aligned}$$

Hereafter, the range of integrations will not be shown and should be clear from the context.

The entropies of order statistics $Y_1, \cdots, Y_n$ are found by noting that $W_i = F_X(Y_i),\ i = 1, \cdots n$. The transformation formula for the entropy applied to $Y_i = F_X^{-1}(W_i)$ gives the following representations of the entropy of order statistics:

$$H(Y_i) = H_n(W_i) - E_{g_i}\left[\log f_X\left(F_X^{-1}(W_i)\right)\right] \tag{2.1}$$

$$= H_n(W_i) - \int f_i(y) \log f_X(y) dy, \tag{2.2}$$

where $H_n(W_i)$ denotes the entropy of the beta distribution shown in (1.2). The expression for beta entropy is

$$H_n(W_i) = \log B(i, n-i+1) - (i-1)[\psi(i) - \psi(n+1)] - (n-i)[\psi(n-i+1) - \psi(n+1)], \tag{2.3}$$

where $\psi(z) = \dfrac{d \log \Gamma(z)}{dz}$ is the digamma function. Noting that $H_n(W_1) = H_n(W_n) = 1 - \log n - \dfrac{1}{n}$, representation (2.2) generalizes proposition 6.1 of Park (1995) for $H(Y_1)$ and $H(Y_n)$.

The following property of the beta entropy is used in the sequel.

$$\begin{aligned} \Delta_n(i) &= H_n(W_i) - H_n(W_{i+1}) \\ &= [\log(n-i) - \psi(n-i)] - [\log i - \psi(i)] < 0 \text{ for } i < \frac{n}{2} \\ &\qquad\qquad\qquad\qquad\qquad\qquad > 0 \text{ for } i > \frac{n}{2}, \end{aligned} \tag{2.4}$$

and for an even n, $\Delta_n(n/2) = 0$. The inequalities in (2.4) are obtained by noting that $\phi(z) = \log z - \psi(z)$ is a decreasing function; $\phi'(z) = \dfrac{1}{z} - \psi'(z) < 0$, where $\psi'(z)$ is the trigamma function (Mitrinovic, 1970, p. 228).

When F_X has a closed form, the entropy of order statistics can be computed using the entropy expression for the beta distribution (2.3) and evaluating the expectation term in (2.1). As an application of the representation (2.1) consider the following example.

Example 2.1 Let X be a random variable having the exponential distribution $F_X(x) = 1 - e^{-\lambda x}$. For computing $H(Y_n)$, we find $F_X^{-1}(w) = -\lambda^{-1} \log(1-w)$ and the expectation term in (2.1),

$$\begin{aligned} E_{g_i}\left[\log f_X\left(F_X^{-1}(W_i)\right)\right] &= E_{g_i}[\log \lambda + \log(1 - W_i)] \\ &= \log \lambda + \psi(n-i+1) - \psi(n+1). \end{aligned} \tag{2.5}$$

For the sample minimum $i = 1$, (2.3) gives $H_n(W_1) = 1 - \log n - \dfrac{1}{n}$. Evaluating (2.5) and noting that $\psi(n+1) = \psi(n) + \dfrac{1}{n}$, (2.1) gives $H(Y_1) = 1 - \log n\lambda$. Thus in this case, (2.1) gives the result in accord with the known fact that the sample minimum has an exponential distribution with parameter $n\lambda$. However, the case of the sample maximum is more complicated. The distribution function of Y_n is $F_n(y) = \left(1 - e^{-\lambda x}\right)^n$ and the density is $f_n(y) = n\lambda \left(1 - e^{-\lambda x}\right)^{n-1} e^{-\lambda x}$. Noting that $H_n(W_n) = 1 - \log n - \dfrac{1}{n}$, the formula (2.1) simply gives $H(Y_n) = 1 - \log n - \log \lambda + \psi(n) + \gamma$, where $\gamma = -\psi(1) = .5772\cdots$ is the Euler constant. Note that $H(Y_n) - H(Y_1) = \psi(n) + \gamma \geq 0$. The equality holds only when $n = 1$. That is uncertainty about the maximum is always more than the minimum in exponential samples. The asymptotic difference is $H(Y_n) - H(Y_1) \approx \log(n) + \gamma$. Finally, it can be shown that for all $i = 1, \cdots, n-1$,

$$H(Y_{i+1}) - H(Y_i) = \frac{1}{n-i} - \Delta_n(i) \geq 0.$$

The inequality can be seen from (3.3) in Section 3. That is the entropy of the ith order statistic of sample from the exponential distribution is increasing in i.

The representation (2.1) also facilitates development of results about the entropy of order statistics. The following theorem provides bounds for the entropy of order statistics $H(Y_i)$ in terms of the entropy of data distribution $H(X)$.

Theorem 2.1. For any random variable X with entropy $H(X) < \infty$ the entropy of order statistics $Y_i,\ i = 1, \cdots, n$ is bounded as follows:

(a) Let B_i denote the ith term of the binomial probability $\text{Bin}(n-1, p_i)$, $p_i = \dfrac{i-1}{n-1}$. Then

$$H_n(W_i) + nB_i[H(X) + I(A)] \leq H(Y_i) \leq H_n(W_i) + nB_i[H(X) + I(\bar{A})],$$

where

$$I(A) = \int_A f(x) \log f(x) dx,$$

$A = \{x : f(x) \leq 1\}$, and $\bar{A} = \{x : f(x) > 1$.

(b) Let $M = f_X(m) < \infty$, where $m = \sup\{x : f_X(x) \leq M\}$ is the mode of the distribution. Then

$$H_n(W_i) - \log M \leq H(Y_i) \leq H_n(W_i) - \log M + nB_i[H(X) + \log M].$$

Proof.

(a) The mode of the beta distribution $g_i(w)$ is p_i. Thus,

$$g_i(w) \leq g_i(p_i) = \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)} p_i^{i-1}(1-p_i)^{n-i} = nB_i. \tag{2.6}$$

Now,

$$\begin{aligned}
-E_{g_i}\left[\log f_X\left(F_X^{-1}(W_i)\right)\right] &= -\int g_i(w) \log f_X\left(F_X^{-1}(w)\right) dw \\
&= -\int_{A_1} g_i(w) \log f_X\left(F_X^{-1}(w)\right) dw \qquad (2.7) \\
&\quad -\int_{\bar{A}_1} g_i(w) \log f_X\left(F_X^{-1}(w)\right) dw \qquad (2.8) \\
&\leq -\int_{A_1} g_i(w) \log f_X\left(F_X^{-1}(w)\right) dw \\
&\leq nB_i \left[-\int_{A_1} \log f_X\left(F_X^{-1}(w)\right) dw\right] \\
&= nB_i \left[-\int_A f_X(x) \log f_X(x) dx\right] \\
&= nB_i \left[H(X) + \int_{\bar{A}} f_X(x) \log f_X(x) dx\right],
\end{aligned}$$

where $A_1 = \left\{w : f_X\left(F_X^{-1}(w)\right) \leq 1\right\}$, and $\bar{A}_1 = \left\{w : f_X\left(F_X^{-1}(w)\right) > 1\right\}\Big\}$. The first inequality is obtained by noting that the integral in (2.8) is nonnegative. The second inequality is obtained using (2.6). The lower bound of $H(Y_i)$ is obtained similarly by noting that the integral in (2.7) is nonpositive.

(b) Let $Z = MX$ and $V_i = MY_i,\ i = 1, \cdots, n$ denote the order statistics of Z. Then $f_Z(z) = \frac{1}{M} f_X\left(\frac{z}{M}\right) \leq 1$ for all z. Noting that $I(A) = -H(Z)$ and $I(\bar{A}) = 0$, from Part (a) we have

$$H_n(W_i) \leq H(V_i) \leq H_n(W_i) + nB_i H(Z).$$

Using $H(Z) = H(X) + \log M$ and $H(V_i) = H(Y_i) + \log M$ gives the result.

The bounds given in Theorem 2.1 are useful when the probability distribution function F_X does not have a closed form, and thus the density of ordered statistics (1.1) and the beta expectation in (2.1) can not be easily evaluated. The entropy expression for many well known distributions is available, and thus the bounds in Theorem 2.1 are easily computable. When the bounds in both parts of Theorem 2.1 can be computed, one may use the maximum of the two lower bounds and the mimimum of the two upper bounds.

Example 2.2 We compute the bounds for the entropies of the sample minimum and maximum for some well known distributions. Noting $H_n(W_1) = H_n(W_n) = 1 - \log n - \frac{1}{n}$ and $B_1 = B_n = 1$, Part (b) of Theorem 2.1 gives the following bounds:

$$1 - \log n - \frac{1}{n} - \log M \leq H(Y_i) \leq 1 - \log n - \frac{1}{n} - \log M + n[H(X) + \log M],\quad i = 1, n. \tag{2.9}$$

(a) For the uniform distribution over the interval $[a, b]$, $M = (b-a)^{-1}$ and $H(X) = \log(b-a)$. Thus, the equalities in (2.9) hold.

(b) For the exponential distribution with parameter λ, $M = \lambda$ and $H(X) = 1 - \log \lambda$. Thus,

$$1 - \log n - \frac{1}{n} - \log \lambda \leq H(Y_i) \leq 1 - \log n - \frac{1}{n} - \log \lambda + n,\quad i = 1, n.$$

As noted before, $H(Y_1) = 1 - \log n\lambda$. Thus, the difference between $H(Y_1)$ and the lower bound is n^{-1}, which vanishes as $n \to \infty$.

(c) The density function of Pareto distribution with parameters α and β is

$$\begin{aligned} f_X(x) &= \frac{\alpha\beta^\alpha}{x^{\alpha+1}} \text{ for } x \geq \beta > 0,\ \alpha > 0, \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Here, $M = \frac{\alpha}{\beta}$ and $H(X) = \log \frac{\beta}{\alpha} + \frac{1}{\alpha} + 1$. Thus,

$$1 - \log n - \frac{1}{n} + \log \frac{\beta}{\alpha} \leq H(Y_i) \leq 1 - \log n - \frac{1}{n} + \log \frac{\beta}{\alpha} + \frac{n}{\alpha} + n,\quad i = 1, n.$$

The distribution of Y_1 is also Pareto with parameters $n\alpha$ and β. Consequently, the difference between $H(Y_1)$ and the lower bound is n^{-1}, which vanishes as $n \to \infty$.

(d) For any normal distribution with variance σ^2 we have $M = (2\pi\sigma^2)^{-1/2}$ and $H(X) = \frac{1}{2} + \frac{1}{2}\log 2\pi\sigma^2$. Thus,

$$1 - \log n - \frac{1}{n} + \frac{1}{2}\log 2\pi\sigma^2 \leq H(Y_1) = H(Y_n) \leq 1 - \log n - \frac{1}{n} + \frac{1}{2}\log 2\pi\sigma^2 + \frac{n}{2}.$$

The equality $H(Y_1) = H(Y_n)$ follows from the result of Wong and Chen (1990) on the symmetry of the entropy of order statistics of symmetric distributions.

Next we provide some results on the entropy of order statistics in terms of ordering properties of distributions. We need the following definitions in which X and Z denote random variables with distribution functions F_X and F_Z, density functions f_X and f_Z, and survival functions $\bar{F}_X(x) = 1 - F_X(x)$ and $\bar{F}_Z(z) = 1 - F_Z(z)$.

Definition 2.1. A nonnegative random variable X is said to have a decreasing (an increasing) failure rate, (DFR (IFR)) if the failure rate (hazard function) $\lambda_X(t) = f_X(t)/\bar{F}_X(t)$ is decreasing (increasing) in $t \geq 0$. Equivalently, if $\bar{F}_X(x+t)/\bar{F}_X(t)$ is increasing (decreasing) in t for all $x \geq 0$.

Definition 2.2. The random variable X is said to be stochastically less than Z, denoted by $X \leq \lim^{st} Z$, if $\bar{F}_X(v) \leq \bar{F}_Z(v)$ for all v.

Definition 2.3. The random variable X is said to be less than Z in dispersion ordering, denoted by $X \leq \lim^{d} Z$, if and only if $F_X^{-1}(u) - F_X^{-1}(v) \leq F_Z^{-1}(u) - F_Z^{-1}(v)$ for all $0 \leq v < u \leq 1$.

Definition 2.4. The random variable X is said to be less than Z in likelihood ratio ordering, denoted by $X \leq \lim^{\ell r} Z$, if $\dfrac{f_X(x)}{f_Z(x)}$ is decreasing in x.

Definition 2.5. The random variable X is said be less than Z in entropy ordering, denoted by $X \leq \lim^{e} Z$, if $H(X) \leq H(Z)$.

It is well known that $X \leq \lim^{d} Z$ implies $X \leq \lim^{st} Z$ (Bickel and Lehmann 1976) and $X \leq \lim^{\ell r} Z$ implies $X \leq \lim^{st} Z$. It is also known that $X \leq \lim^{d} Z$ implies $X \leq \lim^{e} Z$ (Oja 1981).

Theorem 2.2. Let X and Z be two nonnegative random variables. If $Z \leq \lim^{st} X$ and X is DFR, then $Z \leq \lim^{e} X$.

Proof. Let X be DFR with $Z \leq \lim^{st} X$. Then:

$$-H(Z) = \int f_Z(z) \log f_Z(z) dz$$

$$
\begin{aligned}
&\geq \int f_Z(z) \log f_X(z) dz \\
&= \int f_Z(z) \log \lambda_X(z) dz + \int f_Z(z) \log \bar{F}_X(z) dz \\
&\geq \int f_X(z) \log \lambda_X(z) dz + \int f_X(z) \log \bar{F}_X(z) dz \\
&= \int f_X(z) \log f_X(z) dz = -H(X).
\end{aligned}
$$

The first inequality is implied by the nonnegativeness of Kullback-Leibler information between f_Z and f_X. The second inequality is obtained using the following result: $Z \leq \lim^{st} X$ if and only if for any non-increasing function ϕ, $E_{f_Z}[\phi(Z)] > E_{f_X}[\phi(X)]$.

Corollary 2.2. Let X be a nonnegative random variable having a DFR. If $Y_i \leq \lim^{st} X$, then $Y_i \leq \lim^{e} X$.

Example 2.3 It is well known that the sample minimum Y_1 is stochastically dominated by X. Thus for a DFR distribution, $Y_1 \leq \lim^{e} X$. Important examples of DFR distributions are gamma and Weibull distributions with shape parameters less than one, Pareto distribution, and the mixtures of exponential distributions. Theorem 2.2 and Corollary 2.2 apply to these and other DFR distributions. For example, for the Pareto distribution discussed in Example 2.2, $H(X) - H(Y_1) = \log n + \frac{1}{\alpha}\left(1 - \frac{1}{n}\right) \geq 0$, for all $n \geq 1$.

Theorem 2.3. Let X be a random variable and let $Y_i,\ i = 1, \cdots, n$ denote its order statistics.

(a) If $f_X\left(F_X^{-1}(x)\right)$ is non-decreasing in x, then $H(Y_i)$ is decreasing in i for $i < n/2$.
(b) If $f_X\left(F_X^{-1}(x)\right)$ is non-increasing in x, then $H(Y_i)$ is increasing in i for $i > n/2$.

Proof.

(a) Using (2.1), we have

$$H(Y_{i+1}) - H(Y_i) = -\Delta_n(i) + E_{g_i}\left[\log\left(f_X(F_X^{-1}(W_i))\right)\right] - E_{g_{i+1}}\left[\log\left(F_X^{-1}(W_{i+1})\right)\right],$$

where $\Delta_n(i)$ is defined in (2.4).
Since order statistics are stochastically ordered, we have $W_i \leq \lim^{st} W_{i+1}$. Also $W_i \leq \lim^{st} W_{i+1}$ implies that for any non-decreasing function ϕ, $E_{g_i}[\phi(W_i)] < E_{g_{i+1}}[\phi(W_{i+1})]$. Thus $H(Y_{i+1}) - H(Y_i) \leq 0$ and the result follows.
(b) The proof is similar to (a) and is omitted.

As an application of Theorem 2.3 consider the following example.

Example 2.4 Let X be a random variable with the uniform distribution over the unit interval. Noting that $F_X(x) = x,\ F_X^{-1}(x) = x$, and $f\left(F_X^{-1}(x)\right) = 1$, both conditions of Theorem 2.3 are satisfied. Thus, the entropy of the ith order statistic is decreasing in i for $i < n/2$ and is increasing in i for $i > n/2$. This confirms (2.4).

Theorem 2.4. Let X and Z be two random variables and denote their order statistics by Y_i and $V_i,\ i = 1, \cdots, n$, respectively. If $X \leq \lim^d Z$, then $Y_i \leq \lim^e V_i$.

Proof. $X \leq \lim^d Z$, then $Y_i \leq \lim^d V_i$, see Shaked and Shanthikumar (1994), and hence $Y_i \leq \lim^e V_i$.

3 Discrimination Information Function

This section discusses discrimination information between the distributions of order statistics and the data distribution, the discrimination information between the distributions of the order statistics, and the mutual information between consecutive order statistics.

3.1 Discrimination between order statistics and the data distribution

The Kullback-Leibler discrimination information between the distribution of the order statistics f_i and the data distribution f_X, is given by

$$\begin{aligned} K_n(f_i : f_X) = K(g_i : U) &= \int g_i(w) \log g_i(w) dw \\ &= -H_n(W_i), \end{aligned}$$

where g_i is the beta distribution (1.2) and U is the uniform distribution. The first equality follows from $U = F_X(X)$ being a one-to-one transformation and $W_i = F_X(Y_i)$.

Therefore, the discrimination information between the distribution of order statistics and the data distribution is distribution free and is only a function of the sample size and the index i. As a function of i, $K_n(f_i : f_X)$ is decreasing in i for $i < n/2$ and is increasing in i for $i > n/2$. This is seen by noting that

$$K_n(f_{i+1} : f_X) - K_n(f_i : f_X) = \Delta_n(i),$$

where $\Delta_n(i)$ is defined in (2.4). That is, the information discrepancy between the distribution of order statistics and data distribution decreases up to the median and then increases. Thus, amongst the order statistics, the median has the closest distribution to the data distribution.

We also note that $K_n(f_i : f_X) = H(U) - H_n(W_i)$. That is, the discrimination information between the distribution of order statistics and the data distribution is the difference between the maximum entropy and the entropy of beta distribution over the unit interval.

The bounds in Theorem 2.1 provide the following bounds for the sum of the entropy of order statistics $H(Y_i)$ and the relative entropy of order statistics $K_n(f_i : f_X)$ in terms of the entropy of data distribution:

$$nB_i[H(X) + I(A)] \leq H(Y_i) + K_n(f_i : f_X) \leq nB_i[H(X) + I(\bar{A})]$$

and

$$-\log M \leq H(Y_i) + K_n(f_i : f_X) \leq nB_i H(X) + (nB_i - 1)\log M.$$

Next result relates the average information discrepancy between the distribution of the order statistics and the data distribution and the difference between the average entropy of order statistics and the entropy of the data distribution.

Theorem 3.1. Let

$$\bar{K}(f_i : f_X) = \frac{1}{n}\sum_{i=1}^{n} K_n(f_i : f_X), \quad \text{and} \quad \bar{H}(Y) = \frac{1}{n}\sum_{i=1}^{n} H(Y_i)$$

Then,

$$\bar{K}(f_i : f_X) = H(X) - \bar{H}(Y) = C_n, \tag{3.1}$$

where

$$C_n = -\frac{1}{n}\sum_{i=1}^{n} \log B(i, n-i+1) - \frac{n-1}{2}$$

is a constant.

Proof. The first equality in (3.1) is obtained by noting that

$$\begin{aligned}
\sum_{i=1}^{n} K_n(f_i : f_X) &= \sum_{i=1}^{n} \int f_i(y) \log\left(\frac{f_i(y)}{f_X(y)}\right) dy \\
&= \sum_{i=1}^{n} \int f_i(y) \log f_i(y) dy - \sum_{i=1}^{n} \int f_i(y) \log f_X(y) dy \\
&= -\sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} \int g_i(F_x(y)) f_X(y) \log f_X(y) dy \\
&= -\sum_{i=1}^{n} H(Y_i) - \int \sum_{i=1}^{n} n q_{i-1} f_X(y) \log f_X(y) dy \\
&= -\sum_{i=1}^{n} H(Y_i) + nH(X),
\end{aligned}$$

where $q_{i-1},\ i = 1, \cdots, n$ are binomial probabilities, $\text{Bin}(n-1, p),\ p = F_X(x)$. The last equality is noted by $\sum_{i=1}^{n} q_{i-1} = 1$.

The second equality in (3.1) is obtained as follows:

$$\begin{aligned}\sum_{i=1}^{n} K_n(f_i : f_X) &= \sum_{i=1}^{n} \int f_i(y) \log\left(\frac{f_i(y)}{f_X(y)}\right) dy \\ &= -\sum_{i=1}^{n} \log B(i, n-i+1) \\ &\quad + \sum_{i=1}^{n} \int i f_i(y) \log[F_X(y)] dy + \sum_{i=1}^{n} \int (n-i) f_i(y) \log[1 - F_X(y)] dy \qquad (3.2)\end{aligned}$$

Now, letting $j = i - 1$ and $U = F_X(X)$, we obtain

$$\begin{aligned}\sum_{i=1}^{n} \int i f_i(y) \log[F_X(y)] dy &= n \int \sum_{j=0}^{n-1} (n-1) q_j f_X(y) F_X(y) \log[F_X(y)] dy \\ &= n(n-1) E_U(U \log U) \\ &= -\frac{n(n-1)}{4}.\end{aligned}$$

The last equality is obtained by noting that U is uniform over the unit interval and $E_U(U \log U) = [\psi(2) - \psi(3)]/2 = -1/4$. The second sum and integral in (3.2) can be evaluated similarly. Noting that $1 - U$ is also uniform, we obtain $-n(n-1)/4$ for the second sum and integral in (3.2), which completes the proof.

Wong and Chen (1990) proved the second equality in (3.1) through a tedious induction. The probability integral transformation greatly simplifies the proof. By Theorem 2 of Wong and Chen we also conclude that the average information discrepancy between the distribution of the order statistics and the data distribution is increasing in the sample size n.

Remark. The discrimination information between the data distribution and the distributions of order statistics is

$$K_n(f_X : f_i) = \log B(i, n-i+1) + n - 1.$$

In this case,

$$\begin{aligned}K_n(f_X : f_{i+1}) - K_n(f_X : f_i) = \log\left(\frac{i}{n-i}\right) &< 0 \text{ for } i < \frac{n}{2} \\ &> 0 \text{ for } i > \frac{n}{2}.\end{aligned}$$

The average symmetric divergence between the distribution of the order statistics and the data distribution is simply

$$\bar{J}(f_i, f_X) = \bar{K}(f_i : f_X) + \bar{K}(f_X : f_i) = \frac{n-1}{2}.$$

3.2 Discrimination between order statistics

The discrimination information between distributions of ith and jth order statistics is given by

$$K_n(f_i : f_j) = \log \frac{\Gamma(j)\Gamma(n-j+1)}{\Gamma(i)\Gamma(n-i+1)} - (i-j)[\psi(i) - \psi(n-i)] - \frac{i-j}{n-i}.$$

Consequently, we have

$$K_n(f_{i+1} : f_i) = \frac{1}{i} + \Delta_n(i),$$

and

$$K_n(f_i : f_{i+1}) = \frac{1}{n-i} - \Delta_n(i). \tag{3.3}$$

The symmetric divergence is simply

$$J_n(f_{i+1}, f_i) = K_n(f_{i+1} : f_i) + K_n(f_i : f_{i+1}) = \frac{n}{i(n-i)}.$$

It can be shown that all three measures are decreasing for $i \leq (n+1)/2$ and increasing for $i \geq (n+1)/2$. Moreover, the symmetric divergence is symmetric in i and $n-i$. Therefore, the distribution of the consecutive order statistics become closer to each other as they approach the median from either extremes.

Next we give a result on the discrimination information between the order statistics in two samples.

Theorem 3.2. Let X and Z be two random variables and let f_i and ℓ_i denote the densities of their order statistics Y_i and V_i, $i = 1, \cdots, n$, respectively.

(a) If $Z \leq \lim^{st} Y_i$ and $X \leq \lim^{\ell r} V_i$ for an $i < n/2$, then $K_n(f_{i+1} : \ell_{i+1}) \leq K_n(f_i : \ell_i)$.
(b) If $Z \geq \lim^{st} Y_{i+1}$ and $X \geq \lim^{\ell r} V_{i+1}$ for an $i > n/2$, then $K_n(f_{i+1} : \ell_{i+1}) \geq K_n(f_i : \ell_i)$, respectively.

Proof.

(a) Write

$$K_n(f_i : \ell_i) = K_n(f_i : f) + \int f_i(x) \log \frac{f(x)}{\ell_i(x)} dx$$

Therefore,

$$\begin{aligned} K_n(f_{i+1} : \ell_{i+1}) - K_n(f_i : \ell_i) &= \Delta_n(i) + \int f_{i+1}(x) \log \frac{f(x)}{\ell_{i+1}(x)} dx - \int f_i(x) \log \frac{f(x)}{\ell_i(x)} dx \\ &\leq \int f_{i+1}(x) \log \frac{f(x)}{\ell_{i+1}(x)} dx - \int f_i(x) \log \frac{f(x)}{\ell_i(x)} dx \\ &\leq \int f_{i+1}(x) \log \frac{f(x)}{\ell_{i+1}(x)} dx - \int f_{i+1}(x) \log \frac{f(x)}{\ell_i(x)} dx \\ &= \int f_{i+1}(x) \log \frac{\bar{L}(x)}{L(x)} dx \end{aligned}$$

$$\leq \int \ell(x) \log \frac{\bar{L}(x)}{L(x)} dx$$
$$= 0.$$

The first inequality comes from the fact that $\Delta_n(i) < 0$. The second inequality is due to the facts that $Y_i \leq \lim^{st} Y_{i+1}$ and $X \leq \lim^{\ell r} V_i$. Finally, the last inequality comes from the fact that $Z \leq \lim^{st} Y_{i+1}$ and $\bar{L}(x)/L(x)$ is decreasing in x. This completes the proof.

(b) The proof is similar to part (a) and is omitted.

As an application of Theorem 3.2 consider the following example.

Example 3.1 Let X and Z be two random variables having exponential distributions with parameters λ_1 and λ_2. Denote order statistics Y_i and V_i, $i = 1, \cdots, n$ as before. Take $\lambda_2 = n^{-1}\lambda_1$, then it implies that $X \leq \lim^{\ell r} V_1$ and all the assumptions in Theorem 3.2 hold.

3.3 Mutual information between consecutive order statistics

The sequence of order statistics $Y_1, \cdots, Y_n$ has a Markovian property. We can measure the degree of dependency among $Y_1, \cdots, Y_n$ by the mutual information between consecutive order statistics

$$M_n(Y_i, Y_{i+1}) \equiv K_n(f_{i,i+1} : f_i f_{i+1})$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{y_{i+1}} f_{i,i+1}(y_i, y_{i+1}) \log \left(\frac{f_{i,i+1}(y_i, y_{i+1})}{f_i(y_i) f_{i+1}(y_{i+1})} \right) dy_i \; dy_{i+1}$$

where the joint density of (Y_i, Y_{i+1}), $i = 1, \cdots n-1$ is

$$f_{i,i+1}(y_i, y_{i+1}) = \frac{\Gamma(n+1)}{\Gamma(n-i)\Gamma(i)} [F_X(y_i)]^{i-1} [1 - F_X(y_{i+1})]^{n-i-1} f_X(y_i) f_X(y_{i+1}) \text{ for } y_i < y_{i+1},$$
$$= 0 \qquad \text{otherwise.}$$

The next Theorem gives the mutual information $M_n(Y_i, Y_{i+1})$ and its properties.

Theorem 3.3. Let X be a random variable with distribution $f_X(x)$ and let Y_i, $i = 1, \cdots, n$ denote its order statistics.

(a) The mutual information between consecutive order statistics is distribution free and is given by

$$M_n(Y_i, Y_{i+1}) = M_n(W_i, W_{i+1})$$
$$= -\log \binom{n}{i} - i\psi(i) - (n-i)\psi(n-i) + n\psi(n) - 1, \tag{3.4}$$

where W_i, $i = 1, \cdots, n$ are the order statistics of the sample from the uniform distribution.

(b) For a given sample size n, the mutual information between consecutive order statistics is symmetric in i and $n-i$, increases in i for $i < n/2$, and decreases in i for $i > n/2$.
(c) The mutual information $M_n(Y_i, Y_{i+1})$ is increasing in n.

Proof.

(a) The first equality in (3.4) follows from the fact that the mutual information is invariant under one-to-one transformation, $W_i = F_X(Y_i)$. Then the second equality is obtained using the beta marginals (1.2) for i and $i+1$ and the joint density

$$\begin{aligned} g_{i,i+1}(w_i, w_{i+1}) &= \frac{\Gamma(n+1)}{\Gamma(n-i)\Gamma(i)} w_i^{i-1}(1-w_{i+1})^{n-i-1} && \text{for } 0 \le w_i < w_{i+1} \le 1, \\ &= 0 && \text{otherwise.} \end{aligned}$$

(b) The symmetry in i and $n-i$ is clear in the expression (3.4). Taking the derivative with respect to i and using the recurrence formula for the digamma function give

$$M_n'(Y_i, Y_{i+1}) = (n-i)\psi'(n-i+1) - i\psi'(i+1).$$

To show that the derivative is positive for $i < n/2$ and negative for $i > n/2$, it suffices to show that $z\psi'(z+1)$ is increasing in $z \ge 1$. Using the recurrence formula for the trigamma function, we have

$$\begin{aligned} (z+1)\psi'(z+2) &= (z+1)\left[\psi'(z+1) - \frac{1}{(z+1)^2}\right] \\ &\ge z\psi'(z+1). \end{aligned}$$

The inequality is obtained from $\psi'(z+1) \ge \dfrac{1}{(z+1)}$.

(c) Similarly, the derivative with respect to n is $M_n'(Y_i, Y_{i+1}) = n\psi'(n+1) - (n-i)\psi'(n-i+1) > 0$.

It is known that the order statistics are associated. That is for any two monotone functions, $T_1(y_1, y_2, ..., y_n)$ and $T_2(y_1, y_2, ..., y_n)$, we have $COV[T_1(Y_1, Y_2, ..., Y_n), T_2(Y_1, Y_2, ..., Y_n)] \ge 0$, see Barlow and Prochan(1981). The mutual information $M_n(Y_i, Y_{i+1})$ captures the extent of any form of functional dependency between the order statistics, including the linear dependency. The invariance of $M_n(Y_i, Y_{i+1})$ under the integral transformation of the random variable X is particularly important in this context. Data from any arbitrary distribution F_X can be obtained by transforming a sample of uniform data $u_1, \cdots, u_n$ by the inverse transformation $x_i = F_X^{-1}(u_i),\ i = 1, \cdots, n$. The mutual information function $M_n(Y_i, Y_{i+1})$ preserves the dependency structure of the order statistics of the uniform sample under the transformation.

By Part (b) of Theorem 3.2, the dependency between the consecutive order statistics is symmetric about the median where the extent of the dependency is the most. By Part (c), the extent of the dependency between consecutive order statistics increases with the sample size.

4 Spacings

The set of differences between consecutive order statistics, $S_i = Y_i - Y_{i-1},\ i = 2, \cdots n, S_1 = Y_1$, is referred to as sample spacings. Spacings from the uniform and exponential distributions have elegant distributional structures and are usually used as the benchmarks for studying spacings.

If X is a random variable with the uniform distribution over the unit interval $[0, 1]$, then $S_i,\ i = 1, \cdots n$, are all identically distributed as a Beta$(1, n)$ variable W_1 with density $g_1(w)$ shown in (1.2). Thus, $H_n(S_i) = H_n(W_1) = 1 - \log n - \frac{1}{n}$, $K_n(f_{S_i} : f_X) = -H_n(W_1)$, and $K_n(f_{S_i} : f_{S_j}) = 0$ for all $i \neq j$.

For computing the mutual information between S_i and S_j, we use the joint entropy of the pairs of spacings $(S_i, S_j),\ i \neq j$, which are identically distributed and have the following bivariate density,

$$\begin{aligned} f_{S_i,S_j}(s_1, s_2) &= n(n-1)(1 - s_1 - s_2)^{n-2} \text{ for } s_1, s_2 \geq 0,\ s_1 + s_2 \leq 1, \\ &= 0 \qquad\qquad\qquad\qquad\quad \text{otherwise.} \end{aligned}$$

It can be shown that the joint entropy of this bivariate density is

$$H_n(S_i, S_j) = -\log[n(n-1)] + \left(2 - \frac{1}{n}\right)\left(1 - \frac{1}{n-1}\right), \quad \text{for} \quad i \neq j = 1, \cdots, n.$$

Using $M_n(S_i, S_j) = H_n(S_i) + H_n(S_j) - H_n(S_i, S_j)$, we find that the mutual information between any pair of spacings of the samples from the uniform distribution is given by:

$$\begin{aligned} M_n(S_i, S_j) &= \log\left(1 - \frac{1}{n}\right) + \frac{1}{n-1}, \quad \text{for} \quad i \neq j = 1, \cdots, n \\ &= \log(1 + \rho) - \frac{\rho}{1+\rho}, \end{aligned}$$

where $\rho = \rho_{ij} = -1/n$ is the correlation between the uniform spacings. Thus, as $n \to \infty$, the uniform spacings become less dependent as well as less correlated random variables.

If the distribution of X is uniform over interval $[a, b]$, then spacings may be represented as $S_i = (b-a)D_i,\ i = 1, \cdots n$, where D_i are spacings of the sample from the uniform distribution over the unit interval $[0, 1]$. Thus, $H_n(S_i) = H_n(W_1) + \log(b - a)$ and $H_n(S_i, S_j) = H_n(D_i, D_j) + 2\log(b - a)$, but the discrimination information functions and the mutual information remain unchanged.

If X has an exponential distribution with parameter λ, then its spacings may be represented as

$$S_i = \frac{1}{n - i + 1} X_i, \quad i = 1, \cdots, n,$$

where $X_i,\ i = 1, \cdots n$ are independent and identically distributed exponential random variables with density $f_X(x) = \lambda e^{-\lambda x}$. Thus,

$$H(S_i) = H(X) - \log(n - i + 1), \quad i = 1, \cdots, n.$$

That is, for the exponential samples, $S_i \leq \lim^e X$ for all $i = 1, \cdots n$ and $S_i \leq \lim^e S_j$ for $i < j,\ j = 2, \cdots n$. The discrimination information functions $K_n(f_{S_i} : f_X)$ and

$K_n(f_{S_i} : f_{S_j})$ are free of λ and are easily computable. Because of the independence, $M_n(S_i, S_j) = 0$ for all $i \neq j$.

More generally, for any random variable X with failure rate $\lambda_X(t)$, the spacings admit the following representation:

$$S_i = \frac{1}{(n-i+1)\lambda_X(a_i)} Z_i, \tag{4.5}$$

where $Z_1, \cdots, Z_n$ are independent and identically distributed exponential random variables $F_Z(z) = 1 - e^{-z}$ and $T_{i-1} \leq a_i \leq T_i$ with $T_i = \sum_{j=1}^{i} \frac{1}{n-i+1} Z_i, \quad i = 1, \cdots, n$; see Pyke (1965) for details.

The following theorem gives a large sample result for the entropy of spacings.

Theorem 4.1. Let X be DFR with spacings $S_i,\ i = 1, \cdots, n$. Then for large $n,\ H(S_i)$ is increasing in i.

Proof. For large $n,\ a_i \approx \frac{i}{n}$ (Pyke, 1965). Using representation (4.5) with $a_i \approx \frac{i}{n}$, we have

$$H(S_i) = H(Z_i) - \log(n-i+1) - \log \lambda_X \left(\frac{i}{n}\right)$$

The results is implied by the assumption that X is DFR.

This result is applicable to large samples from gamma and Weibull distributions with shape parameters less than one, Pareto distribution, and the mixtures of exponential distributions.

References

Arnold, B.C., Balakrishnan, N., and H.N. Nagaraja (1992), *A first course in order statistics*, New York: John Wiley.

Barlow, R. E. and F. Prochan (1981), *Statistical Theory of Reliability* and Life Testing", Silver Spring, MD: To Begin With.

Bickel, P. J. and E. L. Lehmann (1976), "Descriptive Statistics for Nonparametric Models. III. Dispersion", *The Annals of Statistics*, 4, 1139-1158.

Ebrahimi, N., Pflughoeft, K., and E. S. Soofi (1994), "Two Measures of Sample Entropy," *Statistics and Probability Letters,* 20, 225-234.

Oja, H. (1981), "On Location, Scale, Skewness and Kurtosis of Univariate Distributions", *Scandinavian Journal of Statistics*, 8, 154-168.

Mitrinovic, D.S. (1970), *Analytic Inequalities*, New York: Springer-Verlag.

Park, S. (1996), "Fisher Information on Order Statistics", *Journal of the American Statistical Association*, 91, 385-390.

Park, S. (1995), "The Entropy of Consecutive Order Statistics", *IEEE Trans. on Information Theory*, IT41, 20003-2007.

Pyke, R. (1965), "Spacings" (with Discussions), *Journal of Royal Statistical Society, Ser. B*, 27, 395-449

Shaked, M. and J.G. Shanthikumar (1994), *Stochastic Orders and Their Applications*, New York: Academic Press.

Wong, K.M. and S. Chen (1990), "The Entropy of Ordered Sequences and Order Statistics", *IEEE Trans. on Information Theory*, IT36, 276-284.

Estimation of the Multivariate Normal Mean Under the Extended Reflected Normal Loss function

Towhidi, M. and Behboodian, J.

P11097

Department of Statistics, Shiraz University,Iran.

Abstract. This paper considers simultaneous estimation of multivariate normal mean vector using the extended reflected normal loss function (Spiring [9]). It is shown that the sample mean $\overline{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_p)'$ is admissible when $p \leq 2$, but for $p \geq 3$, we obtain a class of estimators similar to James-Stein estimators which dominate the sample mean in terms of risks.

Keywords. Admissibility, Inadmissibility, James-Stein Estimator, Reflected Normal Loss Function.

1 Introduction

Let $\mathbf{X} = (X_1, \ldots, X_p)'$ be a normal vector with mean vector $\theta = (\theta_1, \ldots, \theta_p)'$ and covariance matrix $\sigma^2\mathbf{I}$, where σ^2 is known. We use the notation
$\mathbf{X} \sim N_p(\theta, \sigma^2\mathbf{I})$, in this article. We consider the simultaneous estimation of $\theta = (\theta_1, \ldots, \theta_p)'$ by using a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from $N_p(\theta, \sigma^2\mathbf{I})$ under the extended reflected normal loss function, given by

$$L(\delta, \theta) = K\left[1 - exp\{-(\delta - \theta)'\Gamma^{-1}(\delta - \theta)\}\right] \tag{1.1}$$

where $K > 0$, Γ is a constant positive definite matrix. In practice the maximum loss can be a function of many things (e.g., Production resources, cost of identification, rework and liabilities) but generally it is finite. As a result the quadratic loss function, with its infinite maximum loss, is often inadequate in describing the loss function associated with a product and has been criticized by some researchers (e.g., Tribus and Szonyi [13], Leon and Wu [8]). The bounded loss function (1.1) was introduced by Spiring [9] for the first time. This loss is a bounded and increasing function of the quadratic loss.

To estimate θ with $N = 1$ and $\sigma = 1$, Stein [10] showed that $\mathbf{X}$ is inadmissible when $p \geq 3$ under squared error loss. James and Stein [7] showed that the following estimator, known as J-S estimator,

$$\delta(\mathbf{X}) = \left(1 - \frac{p-2}{\sum_{i=1}^{p} X_i^2}\right)\mathbf{X}$$

has uniformly smaller risk than $\mathbf{X}$, for all θ. Strawderman [12], Efron and Morris [6], and Casella and Hwang [4] studied the problem of estimating multivariate normal mean vector under quadratic loss function. Brandwein and Strawderman [3] provided minimax estimators for the mean of a spherically symmetric distribution with concave loss. Chung and Kim [5] investigated the admissibility of the sample mean $\bar{\mathbf{X}}$ under balanced loss function. (see Zellner [14])

In section 2 of this paper, using the limiting Bayes method, we show that $\bar{\mathbf{X}}$ is admissible when $p \leq 2$ under the loss (1.1). In section 3, we obtain an estimator similar to J-S estimator under the loss (1.1) when $p \geq 3$, in the following form

$$\delta^*(\bar{\mathbf{X}}) = \left(1 - \frac{c^*}{\bar{\mathbf{X}}'\Gamma^{-1}\bar{\mathbf{X}}}\right)\bar{\mathbf{X}}$$

and we show that δ^* dominates the usual estimator $\bar{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X_i} = (\bar{X}_1, \ldots, \bar{X}_p)'$, where $\mathbf{X_i} = (X_{i1}, \ldots, X_{ip})'$ and $\bar{X}_j = \frac{1}{N}\sum_{i=1}^{N} X_{ij}; j = 1, \ldots, p$.

2 Admissibility of $\bar{X}$ when $p \leq 2$

In this section, we consider the admissibility of $\bar{\mathbf{X}}$ when $p = 1$ and 2. We show that $\bar{\mathbf{X}}$ is admissibile, using the standard Blyth's technique [2].

Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be a random sample from $N_p(\theta, \mathbf{I})$ with the prior normal distribution $\pi_a(\theta)$, where θ has the mean vector zero and covariance matrix $\frac{1}{a}\mathbf{I}$. It is easy to show that the Bayes estimator of θ w.r.t. $\pi_a(\theta)$ under the extended reflected normal loss function is

$$\delta_a(\bar{\mathbf{X}}) = \frac{N\bar{\mathbf{X}}}{N+a}$$

with the risk function,

$$\begin{aligned} R(\theta, \delta_a) &= K\left[1 - E\left[exp\left\{-(\frac{N\bar{\mathbf{X}}}{N+a} - \theta)'\Gamma^{-1}(\frac{N\bar{\mathbf{X}}}{N+a} - \theta)\right\}\right]\right] \\ &= K[1 - (\frac{2\pi}{N})^{-\frac{p}{2}}\int exp\{-(\frac{N\mathbf{x}}{N+a} - \theta)'\Gamma^{-1}(\frac{N\mathbf{x}}{N+a} - \theta) \\ &- \frac{N}{2}(\mathbf{x} - \theta)'(\mathbf{x} - \theta)\}d\mathbf{x}] \end{aligned}$$

Now using the fact that for any matrices C_1 and C_2 of appropriate dimensions,

$$(C_1 + C_2)^{-1} = C_1^{-1} - C_1^{-1}(C_1^{-1} + C_2^{-1})^{-1}C_1^{-1} \tag{2.1}$$

it follows that the risk function of the estimator δ_a is equal to

$$K[1 - (\frac{2\pi}{N})^{-\frac{p}{2}}(\frac{N+a}{N})^p\int exp[-\frac{1}{2}\{(\mathbf{y} - \eta)'(2\Gamma^{-1} + \frac{(N+a)^2}{N}\mathbf{I})(\mathbf{y} - \eta)$$

$$+(\frac{a}{N+a})^2\theta'(\frac{1}{2}\Gamma + \frac{N}{(N+a)^2}\mathbf{I})^{-1}\theta\}]d\mathbf{y}]$$

or

$$K\left[1 - \frac{(N+a)^p}{N^{p/2}}|2\Gamma^{-1} + \frac{(N+a)^2}{N}\mathbf{I}|^{-\frac{1}{2}}exp\left\{-\frac{1}{2}(\frac{a}{N+a})^2\theta'(\frac{1}{2}\Gamma + \frac{N}{(N+a)^2}\mathbf{I})^{-1}\theta\right\}\right] \tag{2.2}$$

where η is a function of θ.

Theorem 2.1: $\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_p)'$ is admissible under the loss (1.1) when $p = 1, 2$, where $\bar{X}_j = \frac{1}{N}\sum_{i=1}^{N} X_{ij}, j = 1, \ldots, p$.
Proof: Suppose $\bar{\mathbf{X}}$ is dominated by some estimator $\delta(\bar{\mathbf{X}})$ of θ. Using the continuity of the risk function in θ for an estimator $\delta(\bar{\mathbf{X}})$, it follows that there exists some $\theta_0, \epsilon > 0$ and $\xi > 0$ such that
$R(\theta, \delta) < R(\theta, \bar{\mathbf{X}}) - \epsilon$ for all $\theta_0 - \xi\mathbf{1} < \theta < \theta_0 + \xi\mathbf{1}$
where $\mathbf{1} = (1, 1, \ldots, 1)'$.

Let r_a, r_a^*, r_a^{**} be defined as follows:
r_a = Bayes risk of the Bayes solution δ_a w.r.t. π_a.
r_a^* = Bayes risk of $\bar{\mathbf{X}}$ w.r.t. π_a.
r_a^{**} = Bayes risk of δ w.r.t. π_a.
Then the difference of Bayes risks of $\bar{\mathbf{X}}$ and δ is

$$\begin{aligned} r_a^* - r_a^{**} &\geq \int_{\theta_0 - \xi\mathbf{1}}^{\theta_0 + \xi\mathbf{1}} \left[R(\theta, \bar{\mathbf{X}}) - R(\theta, \delta)\right] \pi_a(\theta) d\theta \\ &\geq \int_{\theta_0 - \xi\mathbf{1}}^{\theta_0 + \xi\mathbf{1}} \epsilon(2\pi)^{-\frac{p}{2}} |\frac{1}{a}\mathbf{I}|^{-\frac{1}{2}} exp(-\frac{a}{2}\theta'\theta) d\theta \\ &\geq c a^{\frac{p}{2}} \end{aligned}$$

the last inequality holds for all $a < 1$, where c is a positive constant not depending on a.
Also, using (2.2), the difference of Bayes risks of $\bar{\mathbf{X}}$ and δ_a is

$$\begin{aligned} r_a^* - r_a &= K\{\frac{(N+a)^p}{N^{p/2}}[|2\Gamma^{-1} + \frac{(N+a)^2}{N}\mathbf{I}||\frac{a}{(N+a)^2}(\frac{1}{2}\Gamma + \frac{N}{(N+a)^2}\mathbf{I})^{-1} + \mathbf{I}|]^{-\frac{1}{2}} \\ &\quad - N^{\frac{p}{2}}|N\mathbf{I} + 2\Gamma^{-1}|^{-\frac{1}{2}}\} \\ &= K\{\frac{(N+a)^p}{N^{p/2}}|(\frac{a}{N} + 1)(2\Gamma^{-1} + \frac{(N+a)^2}{N}\mathbf{I}) - \frac{a(N+a)^2}{N^2}\mathbf{I}|^{-\frac{1}{2}} \\ &\quad - N^{\frac{p}{2}}|N\mathbf{I} + 2\Gamma^{-1}|^{-\frac{1}{2}}\} \\ &= K\left\{(N+a)^{p/2}|2\Gamma^{-1} + (N+a)\mathbf{I}|^{-\frac{1}{2}} - N^{p/2}|N\mathbf{I} + 2\Gamma^{-1}|^{-\frac{1}{2}}\right\} \end{aligned}$$

The second equality is carried out by using the relation (2.1). It can easily be verified that for $p = 1$, the ratio $\frac{r_a^* - r_a^{**}}{r_a^* - r_a}$ tends to infinity as $a \to 0$ and for $p = 2$, this ratio tends to a positive constant as $a \to 0$. Hence, there exists an $a > 0$ such that $r_a^{**} < r_a$ which contradicts the fact that δ_a is a Bayes solution with respect to π_a. Therefore $\bar{\mathbf{X}}$ is admissible for $p = 1, 2$.

3 Inadmissibility of $\bar{X}$ for $p \geq 3$

In this section, we consider estimation of $\theta = (\theta_1, \ldots, \theta_p)'$ from the model of section 1 under the loss (1.1) and find a class of estimators which have uniformly smaller risk than $\bar{\mathbf{X}}$ for $p \geq 3$.

Lemma 3.1: Let $\mathbf{X} = (X_1, \ldots, X_p)'$ be distributed as $N_p(\theta, \mathbf{I})$. If $h : \Re^p \to \Re$ is an almost differentiable function with $E\|\nabla h(\mathbf{X})\| < \infty$, then

$$E[\nabla h(\mathbf{X})] = E[(\mathbf{X} - \theta)h(\mathbf{X})]$$

, where $\nabla h(\mathbf{x}) = \left(\frac{\partial h(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial h(\mathbf{x})}{\partial x_p}\right)'$.
Proof: See Stein [11].

Theorem 3.1: Let the positive values $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$ be the eigenvalues of the matrix Γ. If the estimator δ^c is defined as

$$\delta^c(\bar{\mathbf{X}}) = \left(1 - \frac{c}{\bar{\mathbf{X}}'\Gamma^{-1}\bar{\mathbf{X}}}\right)\bar{\mathbf{X}}$$

where $0 < c < c^*$,$c^* = 2\left[\sum_{i=2}^{p} \frac{1}{2+N\lambda_i} - \frac{1}{2+N\lambda_1}\right]$,then $\delta^c(\bar{\mathbf{X}})$ dominates $\bar{\mathbf{X}}$ in terms of risks under the extended reflected normal loss function (1.1) for $p \geq 3$, when $c^* > 0$.
Proof: For any estimator $\delta(\bar{\mathbf{X}})$, we define a function g as

$$g(\theta, \delta) = E\left[exp\left\{-(\delta(\bar{\mathbf{X}}) - \theta)'\Gamma^{-1}(\delta(\bar{\mathbf{X}}) - \theta)\right\}\right]$$

and show that for all $\theta, g(\theta, \delta^c) \geq g(\theta, \bar{\mathbf{X}})$. We observe that

$$\begin{aligned} g(\theta, \delta^c) &= E\left[e^{-(\bar{\mathbf{X}}-\theta)'\Gamma^{-1}(\bar{\mathbf{X}}-\theta)} e^{-\frac{c^2}{\bar{\mathbf{X}}'\Gamma^{-1}\bar{\mathbf{X}}} + 2c(\bar{\mathbf{X}}-\theta)'\frac{\Gamma^{-1}\bar{\mathbf{X}}}{\bar{\mathbf{X}}'\Gamma^{-1}\bar{\mathbf{X}}}}\right] \\ &\geq E\left[e^{-(\bar{\mathbf{X}}-\theta)'\Gamma^{-1}(\bar{\mathbf{X}}-\theta)}\left\{1 - \frac{c^2}{\bar{\mathbf{X}}'\Gamma^{-1}\bar{\mathbf{X}}} + 2c(\bar{\mathbf{X}} - \theta)'\frac{\Gamma^{-1}\bar{\mathbf{X}}}{\bar{\mathbf{X}}'\Gamma^{-1}\bar{\mathbf{X}}}\right\}\right] \end{aligned} \tag{3.1}$$

This inequality follows using the fact that
$e^{-x} \geq 1 - x \qquad \forall x \in \Re$
Now by defining $\Sigma^{-1} = 2\Gamma^{-1} + N\mathbf{I}$, $A = [a_{ij}]_{p\times p} = \Sigma^{1/2}\Gamma^{-1}\Sigma^{1/2}$, $\mathbf{Y} = (Y_1, \ldots, Y_p)' = \Sigma^{-\frac{1}{2}}\bar{\mathbf{X}}$ and $\beta = \Sigma^{-\frac{1}{2}}\theta$, the inequality (3.1) reduces to

$$g(\theta, \delta^{\mathbf{c}}) \geq g(\theta, \bar{\mathbf{X}}) - N^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}\left\{E\left[\frac{c^2}{\mathbf{Y}'\mathbf{A}\mathbf{Y}}\right] - 2cE\left[(\mathbf{Y} - \beta)'\frac{\mathbf{A}\mathbf{Y}}{\mathbf{Y}'\mathbf{A}\mathbf{Y}}\right]\right\} \tag{3.2}$$

where $\mathbf{Y}$ is distributed as $N_p(\beta, \mathbf{I})$.
Note that by using lemma 3.1, it follows that

$$\begin{aligned} E\left[(\mathbf{Y} - \beta)'\frac{\mathbf{A}\mathbf{Y}}{\mathbf{Y}'\mathbf{A}\mathbf{Y}}\right] &= E\left[\sum_{i=1}^{p}\frac{\partial}{\partial Y_i}\frac{\sum_{j=1}^{p} a_{ij}Y_j}{\sum_i\sum_j a_{ij}Y_iY_j}\right] \\ &= E\left[\frac{(\sum_i a_{ii})(\sum_i\sum_j a_{ij}Y_iY_j) - 2\sum_i(\sum_j a_{ij}Y_j)^2}{(\sum_i\sum_j a_{ij}Y_iY_j)^2}\right] \\ &= E\left[\frac{tr(\mathbf{A})}{\mathbf{Y}'\mathbf{A}\mathbf{Y}} - \frac{2\mathbf{Y}'\mathbf{A}^2\mathbf{Y}}{(\mathbf{Y}'\mathbf{A}\mathbf{Y})^2}\right] \end{aligned}$$

and
$$-E\left[\frac{c^2}{\mathbf{Y'AY}}\right]+2cE\left[(\mathbf{Y}-\beta)'\frac{\mathbf{AY}}{\mathbf{Y'AY}}\right]$$
$$=E\left\{\frac{\mathbf{Y}'[(-c^2+2ctr(\mathbf{A}))\mathbf{A}-4c\mathbf{A}^2]\mathbf{Y}}{(\mathbf{Y'AY})^2}\right\} \tag{3.3}$$

We know that $\mathbf{A}$ is a positive definite matrix and is diagonable as $\mathbf{U'AU}=\mathbf{T}=diag\{t_1,\ldots,t_p\}$, where the positive values $t_1,\ldots,t_p$ are the eigenvalues of $\mathbf{A}$. Now, we have $\mathbf{U'A^2U}=\mathbf{T^2}=diag\{t_1^2,\ldots,t_p^2\}$ and therefore (3.3) reduces to
$$-E\left[\frac{c^2}{\mathbf{Y'AY}}\right]+2cE\left[(\mathbf{Y}-\beta)'\frac{\mathbf{AY}}{\mathbf{Y'AY}}\right]$$
$$=E\left\{\frac{\mathbf{Y}'\mathbf{U}[(-c^2+2ctr(\mathbf{A}))\mathbf{T}-4c\mathbf{T}^2]\mathbf{U}'\mathbf{Y}}{(\mathbf{Y'AY})^2}\right\}$$

According to (3.2), we complete the proof by showing that the matrix
$$(-c^2+2ctr(\mathbf{A}))\mathbf{T}-4c\mathbf{T}^2=diag\{ct_1(-c+2tr(\mathbf{A})-4t_1),\ldots,ct_p(-c+2tr(\mathbf{A})-4t_p)\} \tag{3.4}$$
is positive definite when $0<c<c^*$.
It can be verified that $t_i=\frac{1}{N\lambda_i+2}; i=1,\ldots,p$, where the values $\lambda_1\leq\lambda_2\leq\ldots\leq\lambda_p$ are the eigenvalues of Γ. Hence, the diagonal elements of the diagonal matrix (3.4) is positive when
$$0<c<2tr(\mathbf{A})-\frac{4}{N\lambda_i+2}\quad i=1,\ldots,p$$
This condition is equivalent to $0<c<c^*$ with $c^*=2\left[\sum_{i=2}^{p}\frac{1}{2+N\lambda_i}-\frac{1}{2+N\lambda_1}\right]$.

Corollary 3.1: Let the estimator $\delta^*(\bar{\mathbf{X}})$ be given as
$$\delta^*(\bar{\mathbf{X}})=\left(1-\frac{p-2}{(N+2)\bar{\mathbf{X}}'\bar{\mathbf{X}}}\right)\bar{\mathbf{X}}$$
Now, $\delta^*(\bar{\mathbf{X}})$ dominates $\bar{\mathbf{X}}$ under the loss function (1.1) with $\Gamma=\mathbf{I}$, for $p>2$. This estimator is similar to J-S estimator.

4 Conclusions

1. Since $\bar{\mathbf{X}}$ is a minimax estimator for the mean vector θ, hence the estimators δ^c, with $0<c<c^*$, provide a class of minimax estimators which are better than $\bar{\mathbf{X}}$.
2. This class of minimax estimators cannot be achieved by the general result of Brandwein and Strawderman [3].

Acknowledgements:
The authors would like to thank the referees for their constructive suggestions and also the Research Council of Shiraz University.

References

Bekker, A. and Roux, J.J.J.(1995). Bayesian multivariate normal analysis with Wishart prior. *Commun. Statist. Theory Meth.*,**24(10)**, 2485-2497.

Blyth, C.R.(1951). On minimax statistical decision procedures and their admissibility. *The Ann. Math. Statist.*,**22**, 22-42.

Brandwein, A.C. and Strawderman, W.E.(1980). Minimax estimation of location parameters for spherically symmetric distributions with concave loss. *The Ann. Statist.*,**8(2)**, 279-284.

Casella, G. and Hwang, J.T.(1982). Limit expression for the risk of James-Stein estimators. *The Canadian J. Statist.*,**10**, 305-309.

Chung, Y. and Kim, C.(1997). Simultaneous estimation of the multivariate normal mean under balanced loss function. *Commun. Statist. Theory Meth.*,**26(7)**, 1599-1611.

Efron, B. and Morris, M.C.(1973). Stein's estimation rule and its competitor-an empirical Bayes approach. *J. American Statist. Assoc.,* **68**, 117-130.

James, W. and Stein, C.(1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. and Prob.*, 361-379.

Leon, R.V. and Wu, C.F.G.(1992). A theory of performance measures in parameter design. *Statist. Sinica*, **2(2)**, 335-357.

Spiring, F.A.(1993). The reflected normal loss function. *The Canadian J. Statist.* **21(3)**, 321-330.

Stein, C.(1955). Inadmissibility of the usual estimator of a multivariate normal distribution. *Proc. Third Berkeley Symp. Statist. and Prob.,***1**, 197-206.

Stein, C.(1981). Estimation of a multivariate normal distribution. *The Ann. Statist.,* **9**, 1135-1151.

Strawderman, W.E.(1971). Proper Bayes minimax estimators of the multivariate normal distribution. *The Ann. Math. Statist.*, **42**, 385-388.

Tribus, M. and Szonyi, G.(1989). An alternate view of the Taguchi approach. *Quality Progress, May,* 46-52.

Zellner, A.(1994). Bayesian and Non-Bayesian estimation using balanced loss functions. *Statistical Decision Theory and Related Topics V, NewYork, Springer-Verlag*, 377-390.

Index