

Editorial

The University of Neuchâtel celebrated Sir David Cox's 80th birthday on 14th July 2004. The week of 14-18 July 2004, brought together over one hundred people consisting of family members, colleagues and friends of David to the city of Neuchâtel from Australia to the United States passing by Asia, Africa and Europe.

Some of the papers presented during the week of celebration were put together by Anthony Davison, Yadolah Dodge and Nanny Wermuth and published by the Oxford University Press. This issue of *Student*, which is in fact two issues, contains other refereed papers, most of which have been presented during the the conferences.

The journal *Student* was launched in 1995 from the Statistics Group of the University of Neuchâtel with the main objective to publish results of graduate student's researches from the Postgrade in Statistics of the University of Neuchâtel and the rest of the world. For over 10 years and publishing 5 volumes of 4 issues in each volume we think the journal *Student* has reached its stated objective.

The journal *Student* was launched with the following aims and scope: "The aim of *Student* is to present the statistical thoughts of graduate students to the community of researchers, teachers and practitioners of statistics. The journal welcomes the submission of articles in all areas of the theory, methodology and applications of statistics and probability, as well as papers on the history of statistics. Articles of an expository or tutorial nature are very much welcomed. *Student* is published twice a year. The annual subscription price of *Student* is SF 25.00, postage included."

The price of CHF 25.00 was to help graduate students pay not so much for a private journal as the price of private journal is unbelievable even for the libraries of even very rich nations. We also distribute free of charge to many countries that could not effort this amount, thanks to the University of Neuchâtel for helping them to have a free access to the journal.

Data and Statistics, a new journal was incorporated in the journal *Student* in 1996 to help the graduate student understand much better their field and observe how other people in the experimental sciences collect their data and analyse them.

The birth of the Bologna Treaty which has for main aim the unification of the bachelor and master programs in Statistics in all European countries, may require a new journal in statistics for graduate students to accompany their stated aims. When in 1989 we created a Master (called then and today Postgrade) in statistics in the University of Neuchâtel, a Master degree in statistics did not exist in any form or shape. After six years in 1995 with so many master theses written, the creation of the journal became a must. We also thought a lot about a name for the journal. Nothing could be more appropriate than *Student*.

The journal when launched had for its preface the following statements:

While passengers on the plane are divided into three classes: first, business and economy, remember there exist travellers who never fly but still manage to complete their journey. Some, although they can afford the plane fare, even first class, do not fly because of their fear of the big machine. Some, ecofreaks and slowcoaches, embark on a twelve train journey simply because they enjoy riding a train and viewing the scenery. It is to those scientists, teachers, graduate students and practitioners who are not yet sitting on the plane but wish to travel, to publish their first article, that our newly established journal *Student* is addressed.

Soon the journal found its place inside the Current Index of Statistics (from volume 21, 1995), and in the Encyclopedia of Statistical Science, Edited by Johnson and Kotz (1998).

We are now in search of a new home for the journal and we hope the 6th volume appears soon enough to accompany the graduate student's papers world wide with a better format and better materials.

We would like to sincerely thank those people and organizations that help us in all these years, specially the referees who helped with their careful readings graduate students how to write their first paper, to the secretaries who handled with great care all the correspondences, Melanie Geiser and Soazig L'helgouac'h and to the former managing editor, Valentine Rousson, and finally the University of Neuchâtel for their financial help through all these years.

This special issue (Volume 5, No. 3 and 4) is also a closing issue, at least for the moment, until the Journal of *Student* finds a new home.

Neuchâtel, August 2006

Yadolah Dodge
Giuseppe Melfi

Editorial

The University of Neuchâtel celebrated Sir David Cox's 80th birthday on 14th July 2004. The week of 14-18 July 2004, brought together over one hundred people consisting of family members, colleagues and friends of David to the city of Neuchâtel from Australia to the United States passing by Asia, Africa and Europe.

Some of the papers presented during the week of celebration were put together by Anthony Davison, Yadolah Dodge and Nanny Wermuth and published by the Oxford University Press. This issue of *Student*, which is in fact two issues, contains other refereed papers, most of which have been presented during the the conferences.

The journal *Student* was launched in 1995 from the Statistics Group of the University of Neuchâtel with the main objective to publish results of graduate student's researches from the Postgrade in Statistics of the University of Neuchâtel and the rest of the world. For over 10 years and publishing 5 volumes of 4 issues in each volume we think the journal *Student* has reached its stated objective.

The journal *Student* was launched with the following aims and scope: "The aim of *Student* is to present the statistical thoughts of graduate students to the community of researchers, teachers and practitioners of statistics. The journal welcomes the submission of articles in all areas of the theory, methodology and applications of statistics and probability, as well as papers on the history of statistics. Articles of an expository or tutorial nature are very much welcomed. *Student* is published twice a year. The annual subscription price of *Student* is SF 25.00, postage included."

The price of CHF 25.00 was to help graduate students pay not so much for a private journal as the price of private journal is unbelievable even for the libraries of even very rich nations. We also distribute free of charge to many countries that could not effort this amount, thanks to the University of Neuchâtel for helping them to have a free access to the journal.

Data and Statistics, a new journal was incorporated in the journal *Student* in 1996 to help the graduate student understand much better their field and observe how other people in the experimental sciences collect their data and analyse them.

The birth of the Bologna Treaty which has for main aim the unification of the bachelor and master programs in Statistics in all European countries, may require a new journal in statistics for graduate students to accompany their stated aims. When in 1989 we created a Master (called then and today Postgrade) in statistics in the University of Neuchâtel, a Master degree in statistics did not exist in any form or shape. After six years in 1995 with so many master theses written, the creation of the journal became a must. We also thought a lot about a name for the journal. Nothing could be more appropriate than *Student*.

The journal when launched had for its preface the following statements:

While passengers on the plane are divided into three classes: first, business and economy, remember there exist travellers who never fly but still manage to complete their journey. Some, although they can afford the plane fare, even first class, do not fly because of their fear of the big machine. Some, ecofreaks and slowcoaches, embark on a twelve train journey simply because they enjoy riding a train and viewing the scenery. It is to those scientists, teachers, graduate students and practitioners who are not yet sitting on the plane but wish to travel, to publish their first article, that our newly established journal *Student* is addressed.

Soon the journal found its place inside the Current Index of Statistics (from volume 21, 1995), and in the Encyclopedia of Statistical Science, Edited by Johnson and Kotz (1998).

We are now in search of a new home for the journal and we hope the 6th volume appears soon enough to accompany the graduate student's papers world wide with a better format and better materials.

We would like to sincerely thank those people and organizations that help us in all these years, specially the referees who helped with their careful readings graduate students how to write their first paper, to the secretaries who handled with great care all the correspondences, Melanie Geiser and Soazig L'helgouac'h and to the former managing editor, Valentine Rousson, and finally the University of Neuchâtel for their financial help through all these years.

This special issue (Volume 5, No. 3 and 4) is also a closing issue, at least for the moment, until the Journal of *Student* finds a new home.

Neuchâtel, August 2006

Yadolah Dodge
Giuseppe Melfi

David R. Cox: A Brief Biography

Yadolah Dodge

*Institut de Statistique
Université de Neuchâtel
CH-2000 Neuchâtel, Switzerland
Email: Yadolah.Dodge@unine.ch*

David Cox started his career as a statistician in 1944 at the age of 20. Here we give a brief history of his scientific life, with extracts from interviews he has given.

David Roxbee Cox was born in Birmingham on July 15, 1924. He studied mathematics at the University of Cambridge and obtained his Ph.D. from the University of Leeds in 1949. He was employed from 1944 to 1946 at the Royal Aircraft Establishment, and from 1946 to 1950 at the Wool Industries Research Association in Leeds. From 1956 to 1966 he was Reader and then Professor of Statistics at Birkbeck College, London. From 1966 to 1988 he was Professor of Statistics at Imperial College, London. In 1988 he became Warden of Nuffield College and a member of the Department of Statistics at Oxford University. He formally retired from these positions on August 1, 1994.

In 1947 he married Joyce Drummond. They have a daughter and three sons.

Among his many honours, Sir David Cox has received numerous honorary doctorates. He is in particular Doctor Honoris Causa of the University of Neuchâtel and is an honorary fellow of Saint John's College, Cambridge, and of the Institute of Actuaries. He has been awarded the Guy medals in Silver (1961) and Gold (1973) of the Royal Statistical Society. He was elected Fellow of the Royal Society of London in 1973, was knighted by Queen Elizabeth II in 1985 and became an Honorary Member of the British Academy in 2000. He is a Foreign Associate of the US National Academy of Sciences. In 1990 he won the Kettering Prize and Gold Medal for Cancer Research.

At the time of writing this, Sir David Cox has written or co-authored 300 papers and books. From 1966 through 1991 he was the editor of *Biometrika*. He has supervised, collaborated with, and encouraged many students, postdoctoral fellows,



Figure 1. David R. Cox

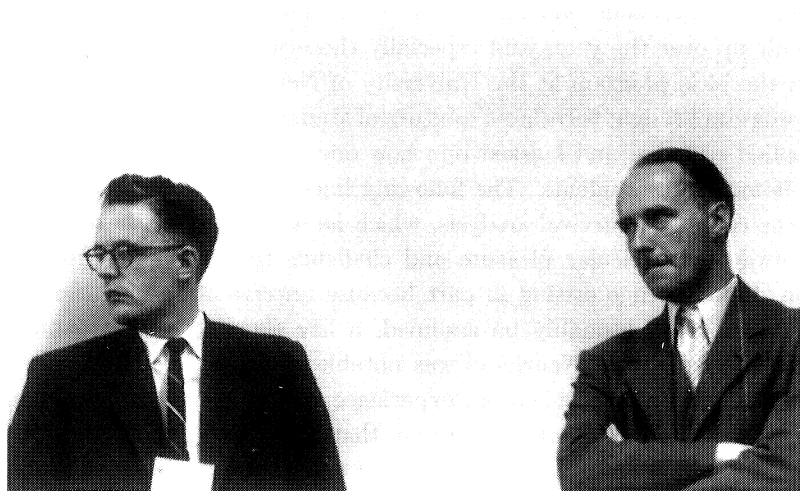


Figure 2. With F.J. Anscombe, Berkeley Symposium, 1960.

and colleagues. He has served as President of the Bernoulli Society, of the Royal Statistical Society, and of the International Statistical Institute. He is now an Honorary Fellow of Nuffield College and a member of the Department of Statistics at the University of Oxford.



Figure 3. D.V. Hinkley, M. Sobel, E.E. Bassett, S. Scriven, T.W. Anderson, P.A. Mills, E.M.L. Beale, A.M. Herberg, E.J. Snell, B.G. Greenberg, D.R. Cox and G.B. Wetherill, London 1988.

He prizes at a personal and scientific level the numerous friendships and collaborations built up over the years and especially those with postgraduate students. He taught on the MSc program at the University of Neuchâtel for over a decade. Each year the program brought together a mixture of students from many different branches of the applied sciences, and I asked him how one should teach statistics to a group of not-so-homogenous students. The following lines are David's random thoughts on teaching the course on survival analysis, which he was giving at the time:

'It is always a particular pleasure and challenge to teach away from one's own institution, the challenge arising in part because uncertainty about the background knowledge that can reasonably be assumed, a key element in the planning of any course. The experience at Neuchâtel was notable especially for the impressive variety of subject fields of specialization, experience and interests of the students. This meant that a key issue was not, as so often, that of providing motivation, but rather that of finding a level of treatment that would enable topics close to the forefront of contemporary research to be addressed without making unreasonable demands on the students theoretical and mathematical knowledge. Of course, it is hard to judge how successfully this was accomplished.

'There are two extreme strategies that may be used in selecting material for such a course. One is to choose, perhaps rather arbitrarily, two or three topics and to develop these carefully in some depth so that hopefully some understanding of principles is achieved. The other, which I find personally more appealing both when I am a student at other people's lectures, and as a teacher, is to range more broadly and to omit detail.

'This raises the following general issue that arises in teaching any topic that is the focus of much research. Our subject has reached the state where in many fields the literature seems overwhelming. In survival analysis not only are there a considerable number of books, some technically very demanding, but also a very large number of research papers, some wholly theoretical, others concerned at least in part with specific applications, but often also containing new methodological points. Further new papers appear virtually daily.

'Yet somehow the implicit message, which I believe to be correct, must be conveyed that this volume of work is not to be thought overwhelming, that it should be regarded as evidence of the openness of the field and that by examining carefully new applications novel work of value can be done. Put differently there are a limited number of key principles and the rest is detail, important in its proper context, but not usually needed. Thus a central issue in each of the specialized areas of statistics is to separate essential principles from detail.

'Much current research in statistics is highly technical; look for example at any issue of the *Annals of Statistics*. It is sometimes argued: how much easier research must have been all those years ago when simple and important results could be obtained by elementary methods.

‘For myself, I very much doubt the correctness of this view, to some extent from personal experience, but more importantly from thinking about the history of key developments in our subject.

‘Take, for example, the Cramér-Rao inequality. A simple elegant very brief proof is available requiring only the Cauchy-Schwarz inequality, itself provable by totally elementary methods. Does this mean that the pioneers had an easy task? It would be interesting to know Professor Rao’s view about this. Certainly the earlier work of Aitken and Silverstone used the mathematically sophisticated ideas of the calculus of variations and was much more difficult.

‘The tentative conclusion I draw is that statistics was and remains a field in which new contributions can be made from all sorts of backgrounds ranging from those with powerful mathematical expertise to those who have new ideas and formulations arising fairly directly out of specific applications. The intellectual vigour of the subject relies on there being a balanced mixture and thus on our subject attracting people with a wide range of interests. One of the strengths of the Neuchâtel course is precisely that.’



Figure 4. With N. Reid, D.A.S. Fraser and J.A. Nelder, Waterloo, 1985.

Another matter of interest for most statisticians is how our leaders think about our field. Here are David Cox’s thoughts:

‘I write some personal thoughts about working as a statistician. I do so largely from a UK perspective and from the viewpoint of a statistician, working mostly but not entirely in a university environment and involved with applications in the natural and social sciences and the associated technologies. The nationality aspect is not very

important but no doubt had I worked in a government statistical office the issues to be discussed would be different, although surely the changes there have been great too.



Figure 5. O.E. Barndorff-Nielsen and D.R. Cox walking and contemplating at Rousham, 1990.

'When I first started statistical work in 1944 there were very few people with a specialized academic training in statistics. Most statistical work was done either by those who like me were trained in mathematics, a few of these with some minimal formal training in statistical ideas, or by natural scientists, especially biologists, or by engineers, who had acquired some knowledge of statistics out of necessity. There were very few books, and those that did exist were mostly accounts of analysis of variance, regression and so on in a form for use by nonspecialists. They gave little or no theory and hence the intellectual basis for tackling new problems was often very unclear. In my second job I was constantly faced with sets of data from carefully designed experiments of not quite standard form and I fear I must have done many analyses that were technically based more on hope obtained by extrapolation from simpler cases than on proper understanding.

'Yet paradoxically by that period, i.e. 1945-1950, many of what one now regards as the standard methods of statistics had been developed, but I suppose were not widely known outside specialized groups. Thus Fisher, and very particularly Yates, had developed design of experiments and analysis of variance into a highly sophisticated art form, least squares and virtually all of its properties having even by then a long



Figure 6. Between A.M. Herzberg and M.S. Bartlett, Oxford, 1990.

history. Principal components dates back at least to the early 1900's and canonical correlation to 1935. Time series methods, both in time and frequency domain had been developed and Neyman had set out definitive ideas on sampling.

'Why were these methods not more widely used? The answer is not so much that they were not known, although this may have been a contributing feature, but much more that some were computationally prohibitive except on a very modest scale. A typical statistician of that period would have the help of at least two or three computers. In those days a computer was a person. Such a person could use an electrically operated calculating machine to do standard balanced analyses of variance, simple regression calculations with up to two or three explanatory variables, simple chi-squared tests and maybe some other more specialized calculations. There was also a rich variety of special types of graph paper for various graphical procedures. Of course other calculations could be done but would need special planning.

'The possibility of numerical errors was ever present. Important calculations had to be checked preferably by a distinct route. I had the enormous good fortune to work for a period with Henry Daniels. He was a very powerful theoretician and I learned pretty well all I knew about theory from him. He was also meticulous. Numerical work had to be done with a sharp pencil entering the answers two (never three) digits per square on quadrille paper (ruled in 1/4 inch squares). Errors had to be rubbed out (never crossed out) and corrected. Tables etc. had to have verbal headings.

'Probably the most distressing calculation was for time series analysis. The data were rounded to two digits and entered twice into a comptometer, an old-style cash register. From this one printed out a paper strip containing the data in columns,

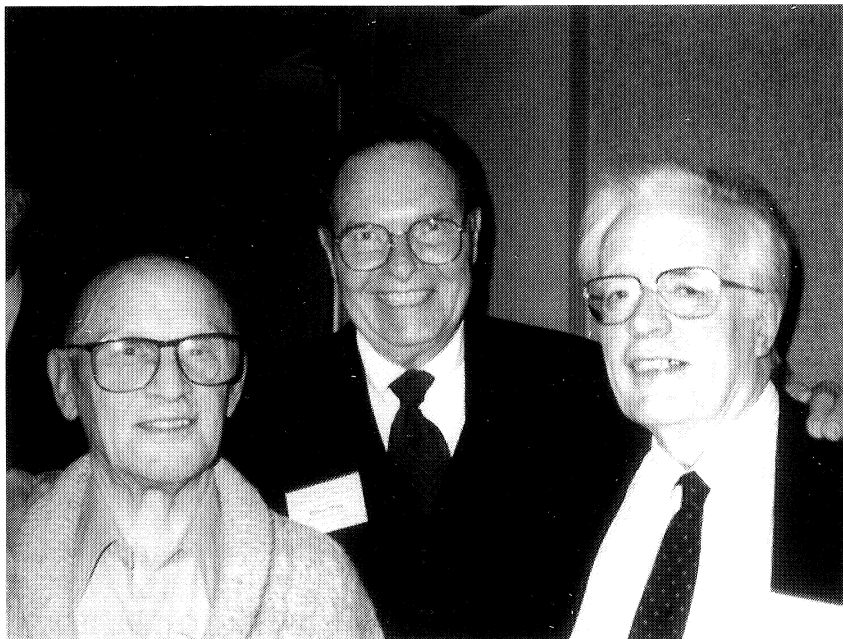


Figure 7. With O. Kempthorne and R. Hogg, Ames, Iowa 1992.

with each value occurring twice. This was cut down the middle to produce two strips. Paper clips were then used to attach these together at displacement h , the desired lag for a serial correlation. Corresponding values were subtracted mentally and then squared mentally and the answer entered, usually by a second person, into an adding machine. This produced $..(y_{i+h} - y_i)^2$, from which the lag h serial correlation can be obtained. This would be repeated for as many values of h as were required, constrained by the need to preserve the sanity of the investigators. Of course I take for granted a knowledge of the squares of the integers up to say 50 or so.

'Matrix inversion takes a time roughly proportional to the cube of the size. The largest I ever did myself was, I think, 7×7 which took, with the obligatory checks, most of one day.

'Now the consequence of all this for the analysis of data was: look very carefully at the data, to assess its broad features, possible defects and so on. Then think. Then compute.

'The idea of an electronic computer was known and I recall in particular the visionary statements of I. J. Good and A. D. Booth, who foresaw the impact computers would have on daily life. It is interesting that it has taken almost 50 years for what they anticipated to come to pass. I suspect that one if not both of them assumed a gentler future in which the miracles made possible by scientific progress would en-

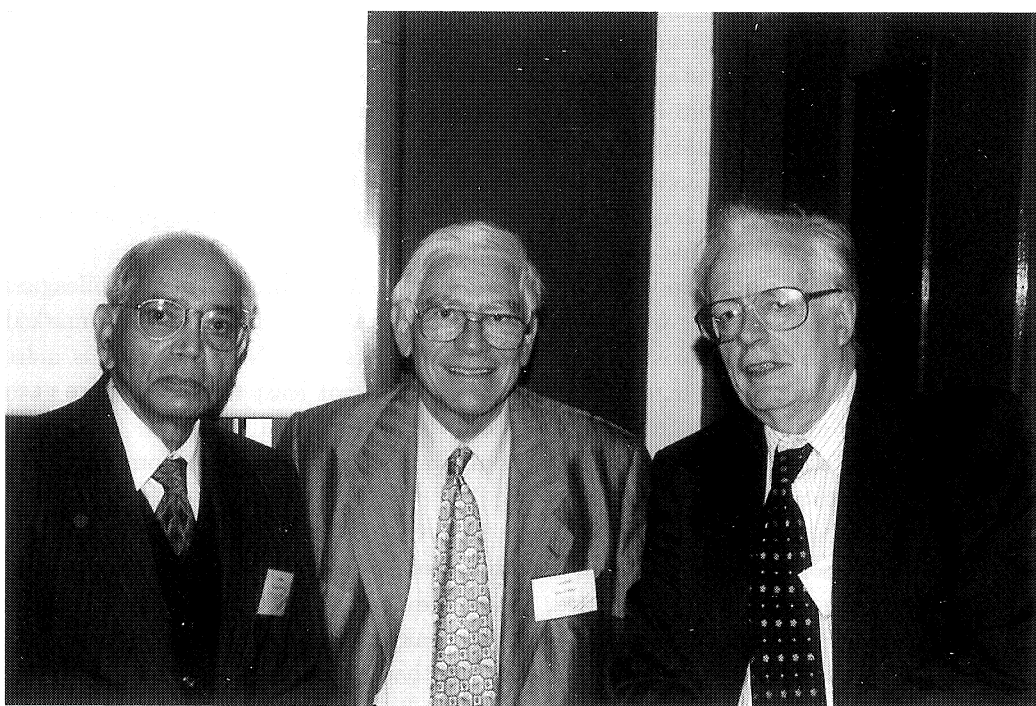


Figure 8. With C.R. Rao and Peter Huber during the Karl Pearson meeting in Paris, May 2000.

hance the general quality of life, not be used as the driving force behind an ever more frantic and in many ways absurd competitive scramble.

'In statistics, as a sweeping oversimplification, the period up to 1975-1980 was taken with getting the classical procedures, and their direct developments, into a form in which they could be relatively painlessly used. Since then the new horizons opened by massive computer power, including the ability to handle large amounts of data, are being explored.

'The consequence of all this for the analysis of relatively modest amounts of data, by contrast with the attitude summarized earlier, is: first compute. Then think.

'In one sense this is a liberation and I do not doubt that on balance it is a good thing, but there is a downside on which we all need to reflect.

'How different now is statistical work from how it was when I started? First there are, of course, many more individuals with statistical training and a much deeper percolation of statistical ideas into subject-matter fields. Think for example of medical research. In many, although certainly not all, areas the specialized subject-matter medical research journals hinge on the use of quite advanced statistical procedures in a way that was unthinkable even 20 years ago.

'The scope for discussion with and help and advice from knowledgeable colleagues is far greater than it was; few statisticians work totally on their own. I remarked above on the virtual nonexistence of books when I started. Now the issue is more the almost overwhelming extent of the literature and not least the appearance each of year of so many books, the great majority of a very good standard. Of course the technical advance of the subject leads to statisticians becoming more and more specialized into particular areas.

'Yet in some sense I wonder if things have changed all that much at a deeper level. Really serious subject-matter research problems virtually always have unique features which demand to be tackled, to some extent at least, from first principles and for this a sound knowledge of the central ideas of our subject and an appropriate attitude to scientific research are required more than very specialist information. My own work has been held back more often by insufficient scientific background in the subject-matter and, especially in my early years, by an inadequate theoretical base in statistics and insufficient mathematical depth, rather than by unfamiliarity with particular specialized techniques.

'Although most research nowadays is done in groups rather than by isolated individuals, research, whether theoretical or applied, remains an intensely personal activity and the best mode of working has to be found by trial and error by each one of us. The subject is at an exciting and demanding phase and I envy those starting on a career in statistical work.'

And finally, here is the humble side of a real scientist, provoked by two questions asked by Nancy Reid in her conversation with David Cox (Reid, 1994):

'The only other thing I wanted to ask you about is something that you mentioned to me in a letter a while ago which I think was connected with your knighthood. Your

39

Effect of serial correlation on maximized log likelihood

Independence model

$$l_{\mathbf{I}}(\theta) = \sum \log f(y_j; \theta)$$

$$U_j = \nabla \log f(Y_j, \theta)$$

Under true model model is OK but there is serial dependence:

$$\text{Cov}(U_j) = i_{\mathbf{I}} = E(-\nabla \nabla^T \log f(Y; \theta))$$

$$\text{Cov}(U_j, U_{j+h}) = \gamma_h i_{\mathbf{I}},$$

where γ_h is $p \times p$ matrix in general.

$$\text{Then } \frac{1}{n} \text{Var}(\sum U_j) \approx i_{\mathbf{I}} + 2 \sum_{h=1}^{\infty} \gamma_h i_{\mathbf{I}}$$

and under true model

$$\sqrt{n}(\hat{\theta}_{\mathbf{I}} - \theta) \text{ is } MN[0, i_{\mathbf{I}}^{-1} \{ i_{\mathbf{I}} + 2 \sum_{h=1}^{\infty} \gamma_h i_{\mathbf{I}} \} i_{\mathbf{I}}^{-1}]$$

i.e.

$$MN[0, i_{\mathbf{I}}^{-1} (I + 2 \sum_{h=1}^{\infty} \gamma_h)]$$

(Cox, 1961;
Agresti,
1983)

Now to assess the quality of the model fit from the independence model we may take $l_{\mathbf{I}}(\theta)$ estimated by $l_{\mathbf{I}}(\hat{\theta}_{\mathbf{I}})$. But

$$l_{\mathbf{I}}(\theta) - l_{\mathbf{I}}(\hat{\theta}_{\mathbf{I}}) \approx -\frac{1}{2} n (\hat{\theta}_{\mathbf{I}} - \theta)^T i_{\mathbf{I}}^{-1} (\hat{\theta}_{\mathbf{I}} - \theta)$$

+ the bias term has expectation

$$-\frac{1}{2} \text{tr} [I + 2 \sum_{h=1}^{\infty} \gamma_h]$$

Figure 9. Sir David Cox's handwriting.

words were roughly that, after feeling when you were younger that you didn't get very much recognition for your work, you now felt you were receiving a "bizarre excess".'

David Cox: 'Yes, I think that sums it up adequately. Well, everybody needs encouragement, and of course as you get older you still need encouragement. But the time you most need it is when you starting. It would be quite wrong to think that people were ever discouraging, they weren't. It was all very low key, typically British understatement sort of thing. You never really knew what people thought, despite the relative frankness of Royal Statistical Society discussions. And I'd published a few papers with very little notion of whether anybody had paid any attention to them until I first went to the United States, where people would come and say, "Oh, I read that paper, I think we could do so and so." That sort of thing is very encouraging. And then, more recently, I've been absurdly lucky with all these pieces of recognition. Of course the system is a bad one in the sense that if you get one piece of recognition it's more likely you'll get another. It ought to be the other way around.'

'Was there a time when you suddenly felt you were a Very Important Person?'

David Cox: 'Well, I hope I've never thought in those terms. In a sense, the only thing that matters is if you can look back when you reach a vast, vast, vast age and say, "Have I done something reasonably in accord with my capability?" If you can say yes, okay. My feeling is in one sense, I've done that: in the tangible sense of books and papers, I've done more than I would have expected. In another sense I feel very dissatisfied: there are all sorts of problems that I nearly solved and gave up, or errors of judgement in doing a little something and not taking it far enough. That I nearly did something you see, this is the irritating thing. You know, if you'd no idea at all, well it doesn't matter, it's irrelevant, but if you feel you were within an inch of doing something and didn't quite do it ...

The Cauchy Distribution

An Enigma with Several Variations

Peter Sprent

*Division of Mathematics
University of Dundee
Scotland
Email: psprent@aol.com*

Abstract : Most statistics students meet the Cauchy distribution as the classic example of a distribution having no finite moments. Unless they later take specialized courses in probability theory or meet it in applications they may learn little more about it. This paper outlines some key properties and mentions some applications mainly in the physical sciences and economics.

Key words: Cauchy distribution, heavy tails, stable distributions, uniform emissions from point sources.

1 Introduction

In undergraduate statistics courses the Cauchy distribution is customarily introduced as the classic example of a distribution that has no finite moments and dismissed on those grounds as a pathological example of little further statistical interest. For many statisticians this, and perhaps a later passing reference to the fact that the t -distribution with 1 degree of freedom is Cauchy distributed, is virtually their only contact with the distribution unless they move on to work in certain specific areas of probability theory or of applied statistics.

It is unfortunate that the *pathological example* tag is freely used even in some well-known text books or in lecturer's handouts where it is common to see it described in terms like '... notable for having no moments and for its frequent appearance in counter examples.' or '... is mostly useful as a bad example.' Jaynes (2003) relates that he was rebuked by a referee for bothering to spend time working on a relevant estima-

tion problem because 'It is well known that the Cauchy distribution is a pathological exceptional case'.

That the distribution is interesting as much more than a pathological example is clear from the excellent account of some of its properties and applications given by Johnson, Kotz and Balakrishnan (1994, chapter 16).

The main fields of application are in the physical sciences, economics and finance.

In this paper I summarize in section 2 a few of the more interesting properties and discuss briefly in section 3 a few applications. Some references to more advanced aspects and to related topics are given in section 4.

2 Some properties of the Cauchy distribution

The general Cauchy distribution, $C(\theta, \lambda)$ has pdf

$$f(x) = \frac{1}{\pi} \frac{\lambda}{\lambda^2 + (x - \theta)^2}$$

cdf

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{x - \theta}{\lambda}$$

and characteristic function

$$\phi(t) = e^{it\theta - |t|\lambda}$$

The special case $C(0, 1)$ is the standard Cauchy distribution. Despite the lack of finite moments the distribution has some elegant properties. The parameters θ and λ have the simple interpretations that the distribution is symmetric about the median θ and the first and third quartiles are at $\theta - \lambda, \theta + \lambda$. Thus λ is the semi-interquartile range. It is easily verified that $f(x)$ has maximum value $1/(\pi\lambda)$ when $x = \theta$ and that $f(\theta - \lambda) = f(\theta + \lambda) = 1/(2\pi\lambda)$ so λ is often described as *the half width at half height*.

Although it was known much earlier that the t -distribution with 1 degree of freedom has a $C(0, 1)$ distribution this is only a special case of a result due to McCullagh (1996) who showed that if X and Y are jointly normally distributed with zero means and covariance matrix

$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix}$$

then $U = X/Y$ is distributed

$$C \left[\frac{\sigma_{xy}}{\sigma_{yy}}, \frac{(\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2)^{1/2}}{\sigma_{yy}} \right]$$

Clearly $1/U = Y/X$ is also Cauchy distributed but the parameters differ. If $\sigma_{xy} = 0$, implying independence of X, Y and if further $\sigma_{xx} = \sigma_{yy}$ then clearly U is distributed $C(0, 1)$, explaining the equivalence for the t -distribution with one degree of freedom. Clearly if X, Y are i.i.d $N(0, \sigma^2)$ implying $U = X/Y$ is $C(0, 1)$ it follows that $1/U$ is also $C(0, 1)$ and by symmetry that $-U$ is $C(0, 1)$ as well, and thus that U and $-1/U$ are identically distributed. Knight (1976) showed that this is a unique property of the $C(0, 1)$ distribution among strictly continuous distributions.

While the heavy-tail characteristic of the Cauchy distribution is well-known it is often not appreciated precisely how the tails compare with those of the normal distribution. It is interesting to compare the normal distribution with zero mean and standard deviation 1.4792 with the $C(0,1)$ distribution. In particular, for each the $\Pr(-1 < X < 1) = 0.05$, while for the normal case $\Pr(-5 < X < 5)$ is 0.9996, whereas for Cauchy $\Pr(-500 < X < 500)$ is only 0.9987.

The Cauchy distribution is stable and infinitely divisible, with the amplitude parameter $\alpha = 1$, while for the normal distribution $\alpha = 2$. For stable distributions $0 < \alpha \leq 2$ and the normal distribution is the only continuous symmetric stable distribution having a finite variance, making it in this context the exceptional case rather than the Cauchy.

For stable distributions α is the fractal or Hausdorff dimension and this governs for example, the self similarity properties associated sums or means of samples of n from iid normal $N(\mu, \sigma^2)$ or Cauchy $C(\theta, \lambda)$ distributions, reflecting the well-known results that the mean in the normal case is distributed $N(\mu, \sigma^2/n)$ and for Cauchy as $C(\theta, \lambda)$. The latter implies, of course, that the mean of a sample of n is no more informative about the location parameter than a single observation, and that the self similarity scaling factor is unity for the sample mean.

The sample median is a possible estimator of θ , or alternatively trimmed or Winsorized means may be used. The efficiency of these estimators is discussed by Johnson, Kotz and Balakrishnan and not surprisingly, in the light of the long tails, severe trimming is needed to get reasonable efficiency.

In view of their stable nature it may be expected that many properties associated with the normal distribution have closely related forms for the Cauchy. For example, if independent $X_i (i = 1, 2, \dots, n)$ are distributed $C(\theta_i, \lambda_i)$ then

$$S = \sum_{i=1}^n X_i$$

is distributed $C(\sum \theta_i, \sum \lambda_i)$ while if X, Y are iid $C(0, 1)$ and a, b are constants then $U = aX + bY$ is $C(0, |a| + |b|)$.

By contrast, the Cauchy distribution has some properties for which there is no normal distribution equivalent. A remarkable characterization due to Williams (1969) is that if $(1 + aX)(a - X)^{-1}$ has the same distribution as X and

$$\frac{1}{\pi} \arctan a$$

is not a rational number then X is $C(0, 1)$. One suspects this result may be of some interest to number theorists. The irrationality condition clearly excludes situations where $\arctan a$ is a rational multiple of π .

While the central limit theorem does not hold for the Cauchy distribution the following elegant result stems from work by Pitman (1961) and other interesting properties of the Cauchy distribution are to be found in Pitman and Williams (1967).

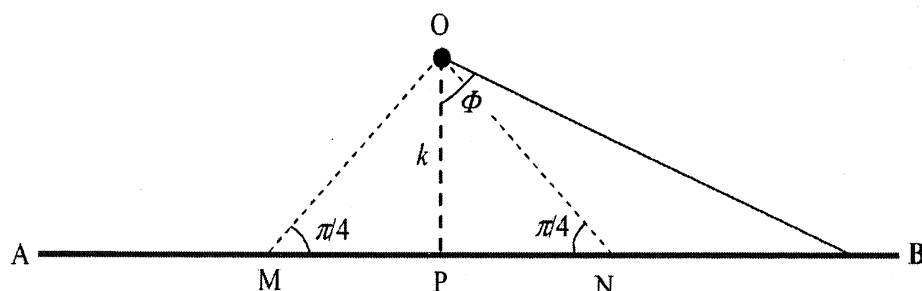


Figure 1. Random emission of particles from a point source O generates $C(0, k)$ distribution of impact points on line at distance k from source.

If $X_1, X_2, \dots, X_n$ are iid with common pdf $f_X(x)$ such that

- (i) $f_X(0) > 0$,
- (ii) $f_X(x)$ is continuous at $x = 0$,
- (iii) $f_X(x)$ has left and right-hand derivatives at $x = 0$,

then $n^{-1} \sum X_i^{-1}$ has a limiting Cauchy distribution as $n \rightarrow \infty$.

Since the reciprocal of a Cauchy distributed variate is also Cauchy distributed it follows that the harmonic mean also has a limiting Cauchy distribution. Pitman often referred to this result as his *eccentric limit theorem*. In particular this result holds if the common distribution is (i) a normal or (ii) a Cauchy distribution.

3 Some applications

Applications of the Cauchy distribution are found mainly but not exclusively in the physical sciences and economics. Several physical applications rely on the fact that if ϕ is uniformly distributed over $(-\pi/2, \pi/2)$ then it is easily established that $X = k \tan \phi$ has a $C(0, k)$ distribution. In Figure 1 if particles are emitted from the point O in directions specified by an angle ϕ where ϕ is distributed uniformly over the interval $(-\pi/2, \pi/2)$ and travel in straight lines, then the distribution of the impact points on a linear receptor AB at distance k from O will be distributed $C(0, k)$.

An equivalent situation is one where the light source in a lighthouse on a rock k miles off a straight shoreline rotates at constant velocity and emits a thin pencil of light and one is interested in the time distribution of the point of impact of the light beam on the shoreline. Referring to Figure I it is clear that the light will impinge on the shoreline AB only when ϕ lies between $-\pi/2$ and $\pi/2$ and the required time

distribution is then $C(0, k)$. A moment's reflection shows that for half the time the beam impinges on the shoreline in the interval MN of length $2k$, consistent with k being the semi-interquartile range, a property mentioned in the previous section.

In the first example above an interesting estimation problem involving the Cauchy distribution arises if the position of O is unknown but a sample of distances of impact points from some arbitrary origin on AB is available. Clearly one then wants good estimators of the median (i.e. effectively an estimate of the point P) in Figure 1 and also of k to estimate the position of O.

Fields in the physical sciences where the Cauchy distribution has been shown to have some relevance include studies of Brownian motion, electrical and magnetic fields produced by auroral electrojets, turbulence, velocity studies in vortex gases, particle path length distributions in streams, earthquake mechanisms and stresses.

The Cauchy distribution is only one of a number of stable distributions associated with various aspects of Brownian motion stemming from work by Lévy dating back to the 1930s, including Lévy (1939, 1965), the latter being an updating of an earlier (1948) publication. In the context of Brownian motion many interesting links between distributions that include the Cauchy are discussed in wide ranging papers by Pitman and Yor (1986(a), 1986(b), 2004). In the last of these they point out at an early stage, for example, that if X has an arc-sine distribution on $(0,1)$ with pdf

$$f_X(x) = \frac{1}{\pi\sqrt{x(1-x)}}, \quad 0 < x < 1$$

then X is distributed as $\cos^2 \theta$ where θ has a uniform distribution over $(0, 2\pi)$ and more interestingly from our point of view that X is also distributed as $1/(1+U^2)$ where U is $C(0,1)$. The former result also holds if X is uniformly distributed over the shorter interval $(0, \pi/2)$.

Some applications involving the Cauchy distribution in some of the physical sciences mentioned above are given by Boteler, Pirjola, and Trichtchenko (2000), Breit and Wigner (1936), Pryce and Ashmore (2003) and Schorghofer and Gille (2002).

An interesting example in economics is given by Phillips (1989) who discusses limited information maximum likelihood (LIML) estimation in linear functional relationships when there are identifiability problems. For a certain carefully specified model he shows that a relevant LIML parameter estimate has a Cauchy distribution and obtains a closely related result for a parameter in a certain time series model where there is also an identifiability problem.

Sir David Cox drew my attention many years ago to papers by Geary (1942, 1943) which implied that identifiability problems were often specifically associated with assumptions of normality. While Phillips uses an approach quite different from that taken by Geary, his results show that while dropping of normality assumptions may ease identifiability problems there is still a strong degree of uncertainty reflected in the long tails and infinite variance associated with the Cauchy distribution.

Other fields of application range from the modeling of stock exchange behaviour to risk analysis and to pattern recognition problems and to numerical quadrature.

Papers by Shlyakhter (1994), by Chang, Kadaba, Doerschuk and Gelfand (2001), by Hamza and Krim (2001), by Sebe, Lew, Cohen, Garg and Huang (2002), by Kim and Seo (1999) cover some topics in these areas.

4 Advanced aspects and some related topics

Brief mention has been made to estimation of parameters in connection with random emission of particles from a point source. There are clearly other contexts where samples of observations may be supposed to be distributed $C(\theta, \lambda)$ and one needs estimates of the parameters. A related problem is to test for goodness of fit when θ and λ are prespecified. There are many papers dealing with various aspects of these problems from both a frequentist and a Bayesian approach.

The nonexistence of moments means that for estimation of the location parameter attention centers on estimators such as the sample median, the mean of the first and third quartiles, a trimmed or Winsorised mean. For the dispersion parameter an estimator such as the sample interquartile range is an obvious candidate.

As already mentioned, the extremely heavy tails associated with the Cauchy distribution mean that where the trimmed mean is used as a location estimator severe trimming is needed to obtain reasonably efficient estimators.

For tests of goodness of fit the long tails often mean that standard tests such as Kolmogorov-Smirnov may be less powerful than tests designed specifically for the Cauchy distribution.

Papers dealing with estimation problems include McCullagh (1996), Hanson and Wolf (1996), Dahiya, Staneski and Chaganty (2001), while goodness of fit tests are discussed by Gürtler and Henze (2000), Matsui and Takemura (2003) and by Rublik (2001, 2003).

The truncated Cauchy distribution is of some theoretical interest insofar as it has finite moments, but is probably of less practical importance than the wrapped Cauchy distribution which features prominently in the analysis of circular or angular data as an alternative to the wrapped normal or to the von Mises distribution. Details are given in Fisher (1993) and in Mardia and Jupp (1999).

In stochastic models Cauchy processes are of interest in several physical contexts in connection with Brownian motion and with Lévy flights. Other extensions related to the basic Cauchy distribution that are not discussed in this paper are to the multivariate Cauchy distribution and to skew Cauchy distributions.

A few of the references cited in this paper are to items that may not be readily available in libraries but in most cases abstracts or even full versions of these are available on the internet and can be accessed by entering an author's name or the title of the paper in a search engine.

References

- [1] Boteler, D.H. Pirjola, R. and Trichtchenko, L. (2000) On calculating the magnetic and electric fields produced at the earth's surface by a "wide" electrojet. *Journal of Atmospheric and Solar-Terrestrial Physics*, **62**, 1311–1315.
- [2] Briet, G. and Wigner, E. (1936) Capture of slow neutrons. *Physical Review* **49**, 519–531.
- [3] Chang, Y-C., Kadaba, S.R., Doerschuk, P.C. and Gelfand, S.B. (2001) Image restoration using recursive Markov random field models driven by Cauchy distributed noise. *Signal Processing Letters, IEEE*, **8**, 65–66.
- [4] Dahiya, R.C., Staneski, P.G. and Chagantry, N.R. (2001) Maximum likelihood estimation of parameters of the truncated Cauchy distribution. *Communications in Statistics: Theory and Methods*, **30**, 1737–1750.
- [5] Fisher, N.I. (1993) *Statistical Analysis of Circular data*. Cambridge University Press.
- [6] Geary, R.C. (1942) Inherent relations between random variables. *Proceedings Royal Irish Academy, A*, **47**, 63–76.
- [7] Geary, R.C. (1943) Relations between statistics: The general and the sampling problem when the samples are large. *Proceedings Royal Irish Academy, A*, **49**, 177–196.
- [8] Gürtler, N. and Henze, N. (2000) Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function. *Annals of Institute of Statistical Mathematics*, **52**, 267–286.
- [9] Hanson, K.M. and Wolf, D.R. (1996) Estimators for the Cauchy Distribution. In *Maximum Entropy and Bayesian Methods*, ed. G.R. Heidbreder, Dordrecht:Kluwer Academic, 255–263.
- [10] Hamza, A.B. and Krim, H. (2001) Image denoising: a nonlinear robust statistical approach. *IEEE transactions on Signal Processing*, **49**, 3045–3054.
- [11] Jaynes, E.T. (2003) *Probability Theory — The Logic of Science*. Cambridge University Press.
- [12] Johnson, N.L., Kotz, S and Balakrishnan, N. (1994) *Continuous Univariate Distributions, Vol 1*. New York, Wiley.
- [13] Kim, J. and Seo, S. (1999) Fourier transformation can improve quadrature efficiency of Laplace distribution. Published online at *InterStat*.
- [14] Knight, F.B. (1976) A characterization of the Cauchy type. *Proceedings of the American Mathematical Society*, **55**, 130–135.
- [15] Lévy, P. (1939) Sur certains processus stochastiques homogènes. *Composito Mathematica*, **7**, 283–339.
- [16] Lévy, P. (1965) *Processus Stochastiques et Mouvement Brownian*. Paris, Gauthier-Villars.
- [17] McCullagh, P. (1996) Möbius transformation and Cauchy parameter estimation. *Annals of Statistics*, **24**, 787–808.
- [18] Mardia, K.V. and Jupp, P.E. (1999) *Directional Data*. Chichester:Wiley.
- [19] Matsui, M. and Takemura, A. (2003). Empirical characteristic function approach for goodness-of-fit tests for the Cauchy distribution with parameters estimated by MLE or EISE. Discussion paper under CIRJE research project *Statistical Aspects*

- of Insurance and Finance.*
- [20] Phillips, P.C.B. (1989) A Little Magic with the Cauchy Distribution. *Cowles Foundation Discussion Papers*, **886**, Cowles Foundation for Research in Economics, Yale University.
 - [21] Pitman, E.J.G. (1961) Some theorems on characteristic functions of probability distributions. *Proceedings 4th Berkeley Symposium on Mathematical Statistics and Probability II*, 393–402.
 - [22] Pitman, E.J.G. and Williams, E.J. (1967) Cauchy-distributed functions of Cauchy variates. *Annals of Mathematical Statistics*, **38**, 916–918.
 - [23] Pitman, J. and Yor, M. (1986a) Asymptotic laws of planar Brownian motion. *Annals of Probability* **14**, 733–779.
 - [24] Pitman, J. and Yor, M. (1986b) Level crossings of a Cauchy process. *Annals of Probability* **14**, 780–792.
 - [25] Pitman, J. and Yor, M. (2004) Some properties of the arc-sine law related to its invariance under a family of rational maps. *IMS Lecture Notes Monograph*, **45**, (in press).
 - [26] Pryce, R.A. and Ashmore, P.E. (2003) Particle path length distribution in meandering gravel-bed streams: results from physical models. *Earthsurface Processes and Landforms*, **28**, 951–966.
 - [27] Rubik, F. (2001) A quantile goodness-of-fit test for Cauchy distributions, based on extreme order statistics. *Applications of Mathematics*, **46**, 339–351.
 - [28] Rubik, F. (2003) On testing goodness-of-fit for Cauchy distribution. *Measurement Science Review*, **3**, 49–52.
 - [29] Schorghofer, N. and Gille, S.T. (2002) Statistics of velocity gradients in two-dimensional Navier–Stokes and ocean turbulence. *Physical Review, E*, **65**, 026307.
 - [30] Sebe, N., Lew, M.S., Cohen, I., Garg, A. and Huang, T.S. (2002). Emotion recognition using a Cauchy naïve classifier. *Proceedings 16th International Conference on Pattern Recognition*, Vol.1. Institute of Electrical and Electronic Engineers, 17–20.
 - [31] Shlyakhter, A.I. (1994) Levy stable distributions. A long-tail generalization of the normal distribution. *Society for Risk Analysis, 1994 Annual General Meeting*.
 - [32] Williams, E.J. (1969) Cauchy-distributed functions and a characterization of the Cauchy distribution. *Annals of Mathematical Statistics*, **40**, 1083–1085.

Estimating the Parameters of Rainfall Models Using Maximum Marginal Likelihood

Paul Northrop

*Department of Statistical Science, University College London
Email: northrop@stats.ucl.ac.uk*

Abstract : Models based on stochastic point processes have been used widely to characterise the temporal evolution of rainfall at a single location. Such models have many uses in hydrology. However, making inferences about the parameters of these models is notoriously difficult. We propose a method of inference in which rainfall occurrence and rainfall amounts are separated. The temporal parameters of the model, that is, the parameters which govern the occurrence of rainfall, are estimated based on the marginal likelihood of the binary sequence of rainfall occurrence indicators. The remaining parameters are estimated using moment methods. The procedure is illustrated by fitting a rainfall model to data from south west England.

1 Introduction

Rainfall is the primary input for most hydrological systems and therefore models of point rainfall time series have potential application to many hydrological problems. For example, one approach to evaluating flood risk routes synthetic rainfall sequences generated from a rainfall model through a rainfall-runoff model to obtain simulated flood responses [16] .

There is an extensive literature on models for rainfall based on stochastic point processes. An early instance is [10] in which rainfall amounts fall instantaneously in time, that is, rain cells have negligible duration. In the single-site model of [14] rain cells are represented as rectangular pulses of rainfall with random duration. This

generalisation enables better representation of rainfall properties over a range of time scales.

Although the models of Rodriguez Iturbe *et al.* [14] have been developed further, by, for example, Rodriguez Iturbe *et al.* [15], Cowpertwait [3], Kakou [9] and Northrop and Stone [12], all single-site rainfall models in this class have the following structure:

1. rain cell origins arrive in a point process in time;
2. each cell deposits rainfall for a random time D , its *duration*, after its origin;
3. the total intensity $Y(t)$ at time t is the sum of all contributions from all cells alive at t .

Most models are based on the assumption that rain cells deposit rainfall at a random intensity X which is constant over its duration - a rectangular pulse of rainfall. It is common to assume that rain cells are mutually independent in all respects and are independent of the process of cell origins. In most models it is assumed that the intensity of each cell is independent of its duration (exceptions are [3] and [9]).

It is common, for reasons of mathematical tractability, to assume that D is exponentially distributed. In most applications it is necessary for the point process in 1. to be clustered so that rain cells cluster to form storms. Alternative rain cell intensity profiles can be accommodated, see, for example, [12].

Much attention has been devoted to theoretical development of more realistic models of this type. However, making inferences about the parameters of these models is notoriously difficult. The most widely-used approach, based on a generalised method of moments, has a number of disadvantages, including the subjectivity involved in choosing the properties used to fit the model. Formulating the likelihood of the full dataset under these models is prohibitive. A more viable alternative is to construct a likelihood based on certain aspects of the data only.

We propose a method of inference in which rainfall occurrence and rainfall amounts are separated. The temporal parameters of the model, that is, the parameters which govern the occurrence of rainfall, are estimated based on the marginal likelihood of the binary sequence of rainfall occurrence indicators. The remaining parameters are estimated using moment methods. We fit a simple rainfall model, based on a Poisson process of rain cell origins, to rainfall data from the HYREX raingauge network [11] in the catchment of the river Brue in south west England.

2 Statistical inference

Point process-based rainfall models are fitted almost exclusively using a generalised method of moments approach in which the parameters of the model are chosen to reproduce selected properties of the data as closely as possible. This may take the form

of selecting as many properties as parameters, seeking to reproduce all these properties exactly, or selecting a larger number of properties and minimizing some measure of the discrepancy between model and data, for example, a sum of (weighted) squared differences. This method has some disadvantages: the choice of fitting properties is subjective, the resulting parameters estimates can be very sensitive to the properties chosen, there may be quite different set of parameter estimates that give similar fits to the properties chosen and inefficient use is made of the data.

Maximum likelihood estimation would provide a more formal method of inference and, via standard asymptotic theory, a means of quantifying parameter uncertainty and comparing the fit of different models. However, the physically appealing short-term dependencies in the model mean that obtaining the likelihood of a time series $Y_1^{(h)}, \dots, Y_n^{(h)}$ of h -hourly rainfall totals, is extremely difficult. The likelihood of the full dataset is the joint distribution of $Y_1^{(h)}, \dots, Y_n^{(h)}$ under the model. Even under the simplest rectangular pulse model the marginal distribution of a single total is very complicated and successive totals have a complex dependency structure.

One way of proceeding within a likelihood framework is to base inference on certain transformations of the original data or on a subset of the data. The former approach is taken by Chandler [2] who constructs an approximate quasi-likelihood based on the Fourier transform of the data. The simplification that enables the likelihood to be constructed results from the approximate large sample normality of the sample Fourier coefficients. This method is, however, based entirely on the mean and the second-order properties of the data and has been found to reproduce rainfall occurrence properties poorly [17].

In this paper we construct a likelihood based on a subset of the full dataset, namely a series of rainfall occurrence indicators.

3 Marginal likelihood based on rainfall occurrences

We construct a binary time series $Z_1^{(h)}, \dots, Z_n^{(h)}$ of rainfall occurrence indicators by defining

$$Z_i^{(h)} = \begin{cases} 0 & \text{if } Y_i^{(h)} = 0, \\ 1 & \text{if } Y_i^{(h)} > 0, \end{cases}$$

where $Y_i^{(h)}$ is the rainfall total in the time interval $((i-1)h, ih)$. Let θ denote the temporal parameters of the model, that is, the parameters which govern the arrival and duration of rain cells, and let ψ denote the parameter(s) of the distribution of X , the 'intensity' parameters. The full likelihood

$$L(\theta, \psi; \mathbf{y}) = P(\mathbf{y}; \theta, \psi) = P(Y_1^{(h)} = y_1, \dots, Y_n^{(h)} = y_n; \theta, \psi)$$

can be factorised as

$$P(\mathbf{y}; \theta, \psi) = P(\mathbf{z}; \theta)P(\mathbf{y}' | \mathbf{z}; \theta, \psi), \quad (1)$$

where $\mathbf{z} = (z_1, \dots, z_n)$ and $\mathbf{y}' = \{y_i | y_i > 0, i = 1, \dots, n'\}$, by partitioning the data into wet/dry indicators and the (ordered) rainfall totals in wet intervals. Clearly

the first term in (1) involves θ but not ψ . The second term depends on both ψ and θ because the duration and location of a rain cell affect the amounts of rainfall deposited into disjoint time intervals.

The general approach we adopt is to base inferences about θ on the *marginal likelihood* $L_T(\theta; \mathbf{z}) = P(\mathbf{z}; \theta)$ and then, conditional on these inferences, to estimate ψ using moment methods. Using $L_T(\theta; \mathbf{z})$ to estimate θ is desirable due to its simplicity relative to $L(\theta, \psi; \mathbf{y})$ and the intuitive separation of rainfall occurrences and rainfall amounts. The efficiency of this approach (relative to the use of $L(\theta, \psi; \mathbf{y})$) depends on the amount of information about θ lost in ignoring the second term in (1). This is difficult to quantify as $P(\mathbf{y}' | \mathbf{z}; \theta, \psi)$ is not available. However, it is not unreasonable to suppose that knowledge of when it rains and when it doesn't may be used as a basis for making inferences about θ . Clearly the level h of temporal aggregation is important - mean cell durations may be expected to be approximately an hour or less, so using daily data is not appropriate.

We note that a similar approach has been described by Guttorp [8] who considers only the occurrence of rainfall, ignoring rainfall amounts. The models considered in [8] are based on cells which deposit rainfall instantaneously in time. Therefore, in contrast with the more realistic models considered here, the temporal and intensity parameters of the model are independently estimable.

In section 4 we derive $L_T(\theta; \mathbf{z})$ for a simple model in which rain cells arrive in a Poisson process. Under this model $L_T(\theta; \mathbf{z})$ has a relatively simple structure. We note that $L_T(\theta; \mathbf{z})$ does not depend on the rain cell intensity profile and therefore may be used to estimate θ regardless of the type of rain cell used in a model.

4 Poisson models

Under a Poisson model rain cell origins arrive in a Poisson process of rate λ per hour and rain cell durations are exponentially distributed with mean μ_D hours. Thus $\theta = (\lambda, \mu_D)$. The number $N(t)$ of rain cells alive at time t follows an immigration-death process [6]. It follows that the number of cells alive at an arbitrary point in time has a Poisson distribution with mean $\lambda\mu_D$ and that

$$q_t = P(N(th) = 0 | N(0) = 0) = \exp\{-\lambda\mu_D[1 - \exp(-th/\mu_D)]\}.$$

The process has a regeneration point at 0, that is, if we observe $Z_i^{(h)} = 0$, there are no cells alive at time ih and, conditional on $Z_i^{(h)} = 0$, $Z_{i+1}^{(h)}, Z_{i+2}^{(h)}, \dots$ are independent of $Z_1^{(h)}, \dots, Z_{i-1}^{(h)}$. Therefore, for the purposes of deriving the likelihood $L_T(\theta; \mathbf{z})$, we can split $\mathbf{z}$ into sequences of wet and dry periods. A wet period of length k is defined as a sequence 1_k of k consecutive positive h -hourly rainfall totals, preceded and followed by zero totals. Thus, the number of zero h -hourly totals and the frequencies of wet periods of different lengths are sufficient statistics for θ under a Poisson model.

Let $w_1, \dots, w_m$ be the lengths of the m wet periods. If $Z_1^{(h)} = 1$, w_1 is censored and if $Z_n^{(h)} = 1$, w_m is censored. We let $I_1 = I(Z_1^{(h)} = 1)$ and $I_m = I(Z_n^{(h)} = 1)$,

where $I(x) = 1$ if x is true and is 0 otherwise. Let m' be the number of dry h -hourly periods preceded by a dry h -hourly period. The fact that the lengths of successive wet periods are independent means that, if $m \geq 1$ and $w_1 \leq n$,

$$L_T(\theta; z) = P(1_{w_1}0)^{I_1} P(01_{w_1}0)^{1-I_1} P(01_{w_m})^{I_m} P(01_{w_m}0)^{1-I_m} \times \\ \times e^{-\lambda h m'} P(0)^{1-m} \prod_{j=2}^{m-1} P(01_{w_j}0), \quad (2)$$

where $P(01_k0) = P(Z_1^{(h)} = 0, Z_2^{(h)} = \dots = Z_{k+1}^{(h)} = 1, Z_{k+2}^{(h)} = 0)$, $P(0) = e^{-\lambda(\mu_D + h)}$ and if $m = 1$ or $m = 2$ the empty product in (2) is 1. If $m = 1$ and $w_1 = n$, that is all periods are wet, $L_T(\theta; z) = P(1_n)$ and if $m = 0$, that is, all periods are dry, $L_T(\theta; z) = \exp\{-\lambda(\mu_D + nh)\}$.

We need to calculate $P(01_k0)$, $P(1_k0)$, $P(01_k)$ and $P(1_k)$, for $k = 1, 2, \dots$. We use the simple relation (noted by Foufoula-Georgiou and Guttorp [7]) that

$$P(x1y) = P(x \cdot y) - P(x0y), \quad (3)$$

where x and y are binary sequences of successive rainfall occurrence indicators and a dot denotes the event '0 or 1'. Repeated application of (3) can be used to express $P(01_k0)$ as a linear combination of probabilities involving only dots and zeros, that is, probabilities that specified intervals of time are dry. Naive use of (3) on a binary sequence with k 1s will result in a linear combination of 2^k probabilities. Computation time can be reduced by defining probabilities such as $P(01_k0)$ inductively. Noting that the $\{Z^{(h)}\}$ process is reversible in time, we obtain

$$P(01_k0) = e^{-\lambda(\mu_D + 2h)} (q_k - e^{-\lambda h} q_{k-1}) - e^{-\lambda h} \sum_{j=1}^{k-1} q_{k-j-1} P(01_j0), \quad k = 1, 2, \dots,$$

where if $k = 1$ the empty sum is zero. Similarly, we obtain

$$P(1_k0) = P(01_k) = e^{-\lambda(\mu_D + h)} (1 - e^{-\lambda h}) - \sum_{j=1}^{k-1} P(01_j0), \quad k = 1, 2, \dots$$

and

$$P(1_k) = 1 - e^{-\lambda(\mu_D + h)} - \sum_{j=1}^{k-1} P(01_j), \quad k = 1, 2, \dots$$

4.1 Estimation of the intensity parameters

Given estimates $\hat{\lambda}$ and $\hat{\mu}_D$ of λ and μ_D we estimate the intensity parameters ψ using moment methods based on the marginal distribution of $Y^{(h)}$. Suppose that X has a distribution indexed by its mean μ_X and its variance σ_X^2 . Rodriguez-Iturbe, Cox, and Isham [14] show that for the Poisson Rectangular Pulses Model (PRPM)

$$E(Y^{(h)}) = h\lambda\mu_D\mu_X$$

and

$$\text{var}(Y^{(h)}) = 2\lambda\mu_D^2(\sigma_X^2 + \mu_X^2) \left[h + \mu_D(e^{-h/\mu_D} - 1) \right].$$

Equating $E(Y^{(h)})$ and $\text{var}(Y^{(h)})$ with their respective sample equivalents, $\bar{Y}^{(h)}$ and $s_{Y^{(h)}}^2$, and solving for μ_X and σ_X^2 gives

$$\widehat{\mu}_X = \bar{Y}^{(h)} / h \widehat{\lambda} \widehat{\mu}_D \quad (4)$$

and

$$\widehat{\sigma}_X^2 = \frac{s_{Y^{(h)}}^2}{2 \widehat{\lambda} \widehat{\mu}_D [h + \widehat{\mu}_D (e^{-h/\widehat{\mu}_D} - 1)]} - \widehat{\mu}_X^2 \quad (5)$$

as moment estimators of μ_X and σ_X^2 . This specifies only the mean and variance of X . Other considerations, such as a simulation-based study of the reproduction of sample extremes, may be used to choose a density for X .

4.2 Fit to the Brue data

We fit the PRPM to data recorded by raingauge 19 in the HYREX network, located at East Hill Lane Fields. The raingauge is of the tipping bucket variety, with a tipping bucket of size 0.2mm. The raw data are a sequence of tip times, each tip indicating that 0.2mm of rainfall has fallen since the previous tip. The raw data are converted into h -hourly rainfall totals, $Y_i^{(h)}$, $i = 1, \dots, n$ by simply counting the number of tips in each time period. Note that if $Y_i^{(h)} > 0$ we are certain that interval i is wet whereas if $Y_i^{(h)} = 0$ it is possible, particularly if $h \ll 1$, that it rained during interval i but sufficiently lightly that the next tip occurred after the end of the interval. It is common to work with hourly data and such data discretisation problems are likely to be negligible at this level of temporal aggregation.

Clearly it is not appropriate to fit the model to the entire dataset as we expect strong seasonal effects. Therefore we fit the model separately to data from each month of the year, viewing data from the same month in different years as independent replications of the rainfall process. In some instances there are periods of several days' duration in which data are missing. We treat the data on either side of the missing periods as independent. Short periods of missing values, if present, could be accommodated by viewing each missing value as a dot and applying (3) as before.

We fit the PRPM to hourly data covering the period 10th October 1993 to 30th September 1998. The maximum marginal likelihood estimates of λ and μ_D are those values which maximise $L_T(\theta, z)$. The intensity parameters μ_X and σ_X^2 are then estimated using (4) and (5). Standard likelihood theory [5] provides a means of calculating standard errors of $\widehat{\lambda}$ and $\widehat{\mu}_D$ and confidence intervals for λ and μ_D . Confidence intervals based on profile-likelihood may be preferable to an interval which is constrained to be symmetric about the estimate.

Figure 1 is a contour plot of $-\log L_T(\theta; z)$ for hourly data from the month of December. The log-likelihood has a clearly defined maximum of -1282.7 at $(\lambda, \mu_D) = (0.083, 1.16)$. Figure 2 shows time series of raingauge data and data simulated from the fitted PRPM. There is a greater tendency for successive peaks to cluster in the raingauge data than in the data simulated from the PRPM.

Seasonal variation of PRPM parameter estimates is illustrated in Figure 3. As

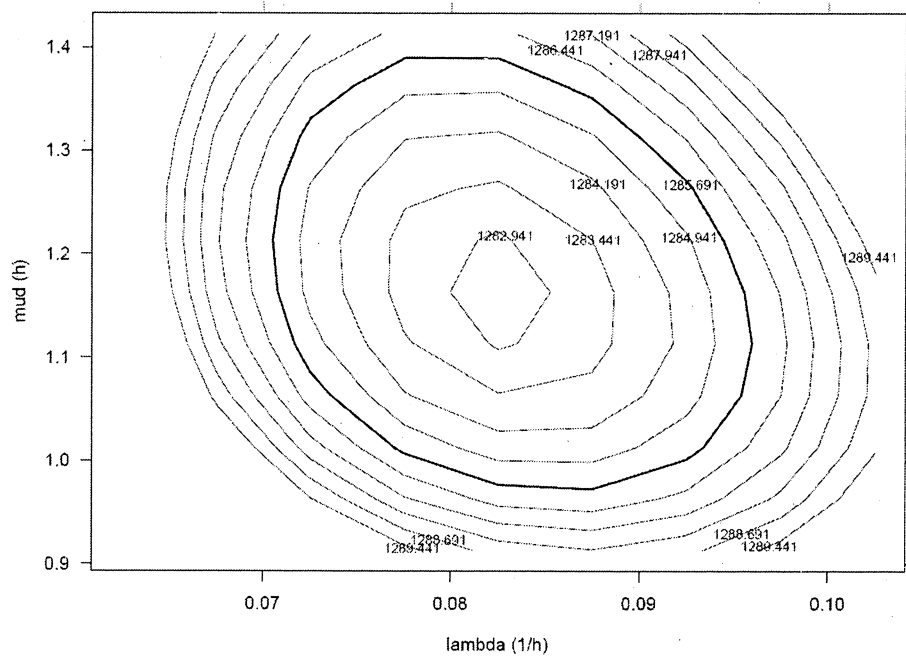


Figure 1. Contour plot of the negative marginal log-likelihood for hourly data from HYREX raingauge 19, December 1993-1997. The bold contour is an approximate 95% confidence region for (λ, μ_D) .

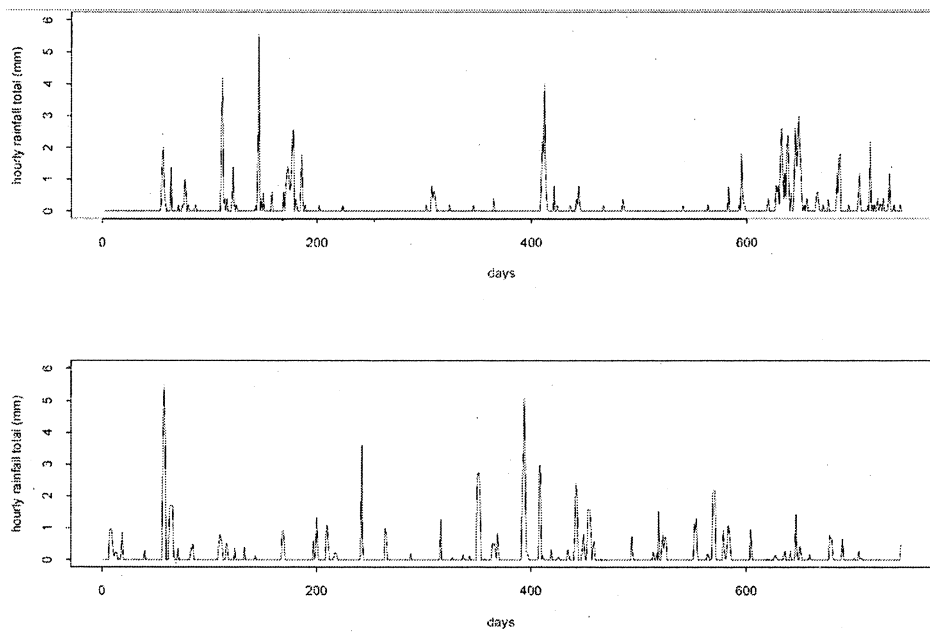


Figure 2. Time series plots of hourly rainfall data. Top: HYREX raingauge 19, December 1994. Bottom: a realisation from the PRPM fitted to hourly data from HYREX raingauge 19, December 1993-1997.

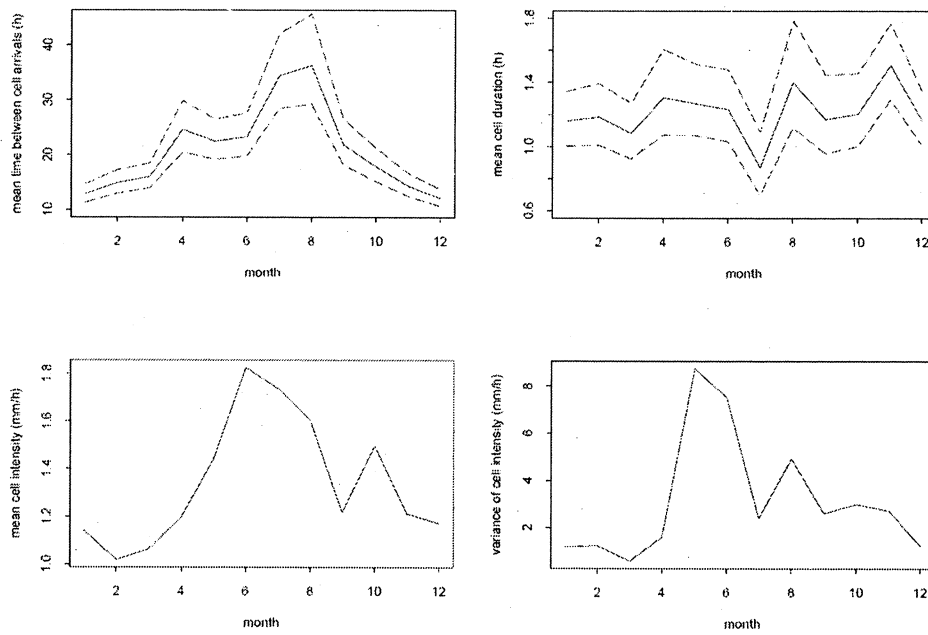


Figure 3. PRPM parameter estimates (top left: $1/\lambda$, top right: μ_D , bottom left: μ_X , bottom right: σ_X^2) against month, with approximate pointwise 95% confidence intervals for $1/\lambda$ and μ_D , constructed based on profile likelihoods.

is expected rain cells tend to be less frequent but more intense in summer than in winter. There is no clear seasonal pattern in the estimates of μ_D .

Inspection of empirical and fitted properties reveal that the PRPM is unable to reproduce simultaneously basic statistical properties of the data from hourly to daily timescales. In particular there is a tendency to underestimate the proportion of dry h -hourly periods for $h > 1$, the degree of underestimation increasing with h , and to overestimate the hourly lag 1 autocorrelation.

5 Concluding remarks

The inadequacy of the fit of the PRPM is to be expected and is not a reflection on the method of estimation. Observational studies, such as [1] and [13], suggest that a model with a clustered point process of rain cells is necessary. [14] developed models based on both Bartlett-Lewis and Neyman-Scott clustering mechanisms in which rain cells cluster to form storms. [15] extend the Bartlett-Lewis model by allowing storms to occur on different timescales. Implementation of marginal likelihood fitting of Bartlett-Lewis models is the subject of ongoing research.

Other extensions are to non-rectangular rain cell profiles, where the marginal likelihood is unaffected by the change in cell shape and cell shape parameter(s) are treated as part of the intensity parameters ψ , and the multi-site models proposed by [4]. A further possibility is to combine the marginal likelihood based on rainfall occurrence with the spectral likelihood of [2], with the aim of constructing a composite likelihood based on rainfall occurrence and on the first and second order properties of rainfall totals.

References

- [1] Austin, P. M. and Houze, R. A. (1972) Analysis of structure of precipitation patterns, *New England. J. Appl. Meteorol.* 11, 926–935.
- [2] Chandler, R. (1997) A spectral method for estimating parameters in rainfall models, *Bernoulli* 3(3), 301–322.
- [3] Cowpertwait, P. S. P. (1994) A generalised point process model for rainfall, *Proc. R. Soc. Lond. A* 447, 23–37.
- [4] Cox, D. and Isham, V. (1994) Stochastic models of precipitation, *Statistics for the Environment* 2, 317–328.
- [5] Cox, D. R. and Hinkley, D. V. (1974) Theoretical statistics. London: Chapman and Hall.
- [6] Cox, D. R. and Miller, H. D. (1965) The theory of stochastic processes. London: Chapman and Hall.
- [7] Foufoula-Georgiou, E. and Guttorp, P. (1986) Compatibility of continuous rainfall occurrence models with discrete rainfall observations, *Wat. Resour. Res.* 22(8), 1316–1322.

- [8] Guttorp, P. (1986) On binary time series obtained from continuous time point processes describing rainfall, *Wat. Resour. Res.* 22(6), 897–904.
- [9] Kakou, A. (1997) Point process based models for rainfall. *Ph.D. thesis*, University of London. Unpublished.
- [10] Le Cam, L. (1961) A stochastic description of precipitation, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 3, ed. J. Neyman, pp. 165–186. Berkeley, California: .
- [11] Moore, R., Jones, D., Cox, D. and Isham, V. (2000) Design of the HYREX rain-gauge network, *Hydrol. Earth. Syst. Sc.* 4(4), 523–530.
- [12] Northrop, P. J. and Stone, T. M. (2005) A point process model for rainfall with truncated Gaussian rain cells. Research report number 251, Department of Statistical Science, University College London. <http://www.ucl.ac.uk/Stats/research/Resrpts/abstracts.html>
- [13] Petterssen, S. (1956) *Weather Analysis and Forecasting*. Volume 2. New York: McGraw Hill.
- [14] Rodriguez-Iturbe, I., Cox, D. R. and Isham, V. (1987) Some models for rainfall based on stochastic point processes, *Proc. R. Soc. Lond. A* 410, 269–288.
- [15] Rodriguez-Iturbe, I., Cox, D. R. and Isham, V. (1988) A point process model for rainfall : further developments. *Proc. R. Soc. Lond. A* 417, 283–298.
- [16] Wheeler, H. S. (2002) Progress in and prospects for fluvial flood modelling, *Phil. Trans. R. Soc. Lond. A* 360(1796), 1409–1431.
- [17] Wheeler, H. S., Isham, V. S., Cox, D. R., Chandler, R. E., Kakou, A., Northrop, P. J., Oh, L., Onof, C. and Rodriguez-Iturbe, I. (2000) Spatial-temporal rainfall fields: modelling and statistical aspects, *Hydrol. Earth. Syst. Sc.* 4(4), 581–601.

Partial Likelihood for Observation Driven Models

Konstantinos Fokianos

Department of Mathematics & Statistics,
University of Cyprus
P.O. Box 20537, 1678 Nicosia, Cyprus,
Email: fokianos@ucy.ac.cy

Abstract : Partial likelihood and observation driven models are two fundamental concepts in modern time series analysis that allow modelling and inference for both qualitative and quantitative time series data. Both of these concepts are developed with emphasis on time series following generalized linear models. The general theory is applied to an autoregressive moving average model for count data.

Key words: log-linear model, random time dependent covariates, binary model, probit model, prediction, residuals.

1 Introduction

Recently, there has been an increasing interest in modelling non standard time series—especially integer valued data—for the purpose of prediction. This article demonstrates the influential work of Professor Cox to this research area by reviewing the *partial likelihood* concept for *observation driven* models. As it will become evident, the combination of these two concepts provides a systematic framework for the analysis of qualitative as well as quantitative time series data. Partial likelihood, introduced by Cox [2, 3], is the proper inferential vehicle since it allows sequential conditional inference with respect to a filtration generated by all the information that is known to the observer at the time of observation. This in turn implies that the data analyst can include autoregressive components, functions of past covariates, all sorts of interactions among covariates, and more generally *time dependent random covariates* in an observation driven model, which is simply a model whose prediction at any time

is determined by the *observations* at previous times, [4]. As a result, estimation, diagnostics, goodness of fit test, prediction are carried out relatively easy by a number of currently available computing environments.

In what follows, the concept of partial likelihood is defined and then the term observation driven model is specialized to time series following generalized linear models. Partial likelihood inference for this special class of time series is discussed briefly—the focus being an autoregressive moving average model for count data—and the article concludes with some remarks. Most of this material is based on the text by Kedem and Fokianos [7] where the interested reader can find more information and applications.

2 On Partial Likelihood

The concept of partial likelihood is better motivated when considering a time series $\{Y_t\}$, $t = 1, \dots, N$, with a joint density $f_{\theta}(y_1, \dots, y_N)$ parameterized by a vector parameter θ . Then, the likelihood is a function of the unknown parameter θ and can be written as

$$f_{\theta}(y_1, \dots, y_N) = f_{\theta}(y_1) \prod_{t=2}^N f_{\theta}(y_t | y_1, y_2, \dots, y_{t-1}). \quad (1)$$

A major problem arises with equation (1) since its generality does not allow inference for the unknown parameter vector. Indeed, if no additional assumptions are made, then as the sample size increases so does the size of θ . Therefore the information obtained concerns an increasing number of parameters, a fact which raises consistency as well as modelling problems. This problem is bypassed by imposing further assumptions which simplify (1). For example, the Markovian property can be employed or, as we shall see shortly, the concept of partial likelihood following [2, 3].

First consider the Markov assumption applied to (1), that is suppose that $\{Y_t\}$, $t = 1, 2, \dots, N$ is a first order Markov process. Then

$$f_{\theta}(y_1, \dots, y_N) = f_{\theta}(y_1) \prod_{t=2}^N f_{\theta}(y_t | y_{t-1}). \quad (2)$$

Since the first factor of the above product does not depend on N , inference regarding θ is carried out using the product term in (2)—this is an example of *conditional likelihood* inference. It can be shown that the logarithm of (2), is a zero mean square integrable martingale, a fact which can be effectively utilized to assess asymptotic properties of the resulting maximum likelihood estimator.

Now, consider the more general concept of *partial likelihood* which was suggested by [3]. The main idea of the method consists of getting a likelihood which is a part of (1). More precisely, consider a time series which is jointly observed with some *random time dependent covariates*, that is we observe (X_t, Y_t) . Then, arguing as before, it can be shown that

$$f_{\theta}(x_1, y_1, \dots, x_N, y_N) = f_{\theta}(x_1) \left[\prod_{t=2}^N f_{\theta}(x_t | d_t) \right] \left[\prod_{t=1}^N f_{\theta}(y_t | c_t) \right] \quad (3)$$

where $d_t = (y_1, x_1, \dots, y_{t-1}, x_{t-1})$ and $c_t = (y_1, x_1, \dots, y_{t-1}, x_{t-1}, x_t)$. The second product on the right hand side of (3) constitutes a partial likelihood according to [3] and can be used for inference. Information about θ is contained in the first product as well, but under some reasonable conditions the loss occurred because of its omission is relatively small. This is precisely the concept of partial likelihood: consider products of conditional densities, where the densities depend on a fixed parameter and where the formed products do not necessarily give complete joint or full likelihood information. The discussion is summarized in the following definition, see Def. 1.1.1 in [7] :

Definition 1 Let $\mathcal{F}_t$, $t = 0, 1, \dots$ be an increasing sequence of σ -fields, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \dots$, and let $Y_1, Y_2, \dots$ be a sequence of random variables on some common probability space such that Y_t is $\mathcal{F}_t$ measurable. Denote the density of Y_t , given $\mathcal{F}_{t-1}$, by $f_t(y_t; \theta)$, where $\theta \in R^p$ is a fixed parameter. The partial likelihood (PL) function relative to θ , $\mathcal{F}_t$, and the data $Y_1, Y_2, \dots, Y_N$, is given by the product

$$PL(\theta; y_1, \dots, y_N) = \prod_{t=1}^N f_t(y_t; \theta) \quad (4)$$

Definition 1, generalizes both concepts of likelihood and conditional likelihood. Notice that partial likelihood becomes an ordinary likelihood when auxiliary information is absent and the data are independent. Furthermore partial likelihood reduces to a conditional likelihood if the covariate process is known throughout the period of observation. The vector θ that maximizes equation (4) is called the maximum partial likelihood estimator. Its theoretical properties, including consistency, asymptotic normality, and efficiency, have been studied extensively in [12] . For further discussion and more references, see [7] .

3 Observation Driven Models for Time Series Following Generalized Linear Models

Suppose that $\{Y_t\}$ is a time series which will be referred as the *response*, and let

$$\mathbf{Z}_{t-1} = (Z_{(t-1)1}, \dots, Z_{(t-1)p})',$$

be the corresponding p -dimensional vector of past explanatory variables or covariates, $t = 1, \dots, N$. Denote by $\mathcal{F}_{t-1}$ the σ -field generated by past values of the response series and past and possibly present values (when known) of the covariates. If $\mu_t = E[Y_t | \mathcal{F}_{t-1}]$ stands for the conditional expectation of the response given the past then the problem is to relate μ_t to the covariates. It turns out that generalized linear models theory as introduced by Nelder and Wedderburn in [10] and further treated by McCullagh and Nelder in [9] provide a unified framework for the analysis of observation driven models for count, binary and more generally categorical data. The definition of such models proceeds as follows:

Consider first the so called random component which postulates that the conditional distribution of the response given the past belongs to the exponential family of distributions, that is for $t = 1, \dots, N$,

$$f(y_t; \theta_t, \phi \mid \mathcal{F}_{t-1}) = \exp \left\{ \frac{y_t \theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t; \phi) \right\}. \quad (5)$$

The parametric function $\alpha_t(\phi)$ is equal to ϕ/ω_t , where ϕ is a dispersion parameter, and ω_t is a known parameter referred to as weight or prior weight. The parameter θ_t is called the *natural* parameter of the distribution. Turning now to the systematic component assume that for $t = 1, \dots, N$, there exists a monotone function $g(\cdot)$ such that,

$$g(\mu_t) = \eta_t = \sum_{j=1}^p \beta_j Z_{(t-1)j} = \mathbf{Z}'_{t-1} \boldsymbol{\beta}. \quad (6)$$

The function $g(\cdot)$ is called the *link function* while η_t is referred to as the linear predictor of the model. Typical choices of the linear predictor include

$$\mathbf{Z}'_{t-1} \boldsymbol{\beta} = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 X_t \sin(\omega_0 t)$$

or

$$\mathbf{Z}'_{t-1} \boldsymbol{\beta} = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-1} X_{t-1} + \beta_4 Y_{t-3} X_{t-1}$$

and so on, given a one dimensional response process $\{Y_t\}$ and a covariate process $\{X_t\}$. Both of the above models are examples of observation driven models since the response process is driven by past values of itself and other covariates while the regression vectors remain constant throughout the observation period.

Following [7] and [6], we obtain

$$\mu_t = E[Y_t \mid \mathcal{F}_{t-1}] = b'(\theta_t).$$

Furthermore,

$$\text{Var}[Y_t \mid \mathcal{F}_{t-1}] = \alpha_t(\phi) b''(\theta_t).$$

The function $V(\mu_t) = b''(\theta_t)$ is called the *variance function* and depends on μ_t . Notice that the link function which is defined by

$$g(\mu_t) = \theta_t(\mu_t) = \eta_t = \mathbf{Z}'_{t-1} \boldsymbol{\beta}$$

is called the *canonical link function*. Certain simplifications occur when the canonical link is entertained.

3.1 Time Series of Counts

Observations of dependent counts can in many cases be modelled successfully through the Poisson distribution. The conditional density of the Poisson distribution with mean μ_t can be written as

$$f(y_t; \theta_t \mid \mathcal{F}_{t-1}) = \exp \{ (y_t \log \mu_t - \mu_t) - \log y_t! \}, \quad t = 1, \dots, N,$$

so that $E[Y_t \mid \mathcal{F}_{t-1}] = \mu_t$, $b(\theta_t) = \mu_t = \exp(\theta_t)$, $V(\mu_t) = \mu_t$, $\phi = 1$, and $\omega_t = 1$ according to (5). The canonical link is

$$g(\mu_t) = \theta_t(\mu_t) = \log \mu_t = \eta_t = \mathbf{Z}'_{t-1} \boldsymbol{\beta}.$$

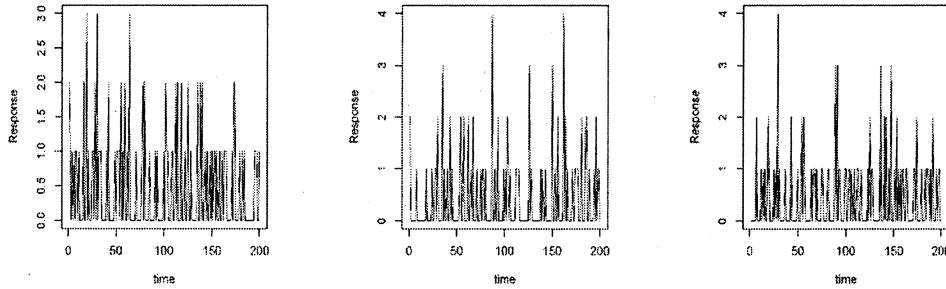


Figure 1. Typical realizations of model (8) with $\beta = (1, -0.5, -0.5)'$ and $N = 200$. From left to right: $\lambda = 0, 0.5, 1$.

Recently Davis *et al.* [5] have provided an example of observation driven models for count data of the following form

$$\mathbf{Z}_{t-1} = (\mathbf{X}'_t, e_{t-1}, \dots, e_{t-p})$$

where $\mathbf{X}_t$ denotes a multivariate auxiliary process and the random sequence $\{e_t\}$ is defined by

$$e_t = \frac{Y_t - \mu_t}{\mu_t^\lambda}, \quad t = 1, \dots, N, \quad (7)$$

for $\lambda \geq 0$. Notice that Li [8], Shepard [11] and Benjamin *et al.* [1] have considered a similar sequence moving average structure but with different fixed values of λ . Hence, the model under consideration can be written as

$$\log \mu_t(\beta) = \mathbf{X}'_t \beta_0 + \sum_{i=1}^p \beta_i e_{t-i}, \quad t = 1, \dots, N$$

where the dimension of the parameter vector β_0 matches the dimension of $\mathbf{X}_t$ and $\beta = (\beta'_0, \beta_1, \dots, \beta_p)'$. For example, suppose that

$$\log \mu_t(\beta) = \beta_{01} + \beta_{02} \cos\left(\frac{2\pi t}{12}\right) + \beta_1 e_{t-1} \quad (8)$$

for $t = 1, \dots, N$ and set $\beta = (\beta_{01}, \beta_{02}, \beta_1)'$ and $\mathbf{Z}_{t-1} = (1, \cos(2\pi t/12), e_{t-1})'$. Figure 1 display realizations of the observed time series of counts for different parameter values of λ showing reduced data oscillation for all λ with most of the observed values falling between the range 0 to 4.

4 Partial Likelihood Inference

We now turn to the partial likelihood inference problem for time series following generalized linear models. Suppose that $\{Y_t\}$, $t = 1, \dots, N$, is a response time series conditionally distributed as (5) and $\{\mathbf{Z}_{t-1}\}$ stands for a p -dimensional vector of ran-

dom time dependent covariates, g is a given link function, and that the dispersion parameter ϕ is known. According to Definition 1 the partial likelihood of the observed series is

$$\text{PL}(\beta) = \prod_{t=1}^N f(y_t; \theta_t, \phi \mid \mathcal{F}_{t-1}).$$

Then from (5), the log-partial likelihood, $l(\beta)$, is given by

$$\begin{aligned} l(\beta) &= \sum_{t=1}^N \log f(y_t; \theta_t, \phi \mid \mathcal{F}_{t-1}) \\ &= \sum_{t=1}^N \left\{ \frac{y_t \theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t, \phi) \right\}. \end{aligned}$$

The partial score is defined as $\nabla l(\beta)$, the first derivative of the log-partial likelihood function with respect to the vector of unknown parameters β . Straightforward calculation shows that it is a p -dimensional vector given by

$$\mathbf{S}_N(\beta) \equiv \nabla l(\beta) = \sum_{t=1}^N \mathbf{Z}_{t-1} \frac{\partial \mu_t(Y_t - \mu_t(\beta))}{\partial \eta_t} \frac{1}{\sigma_t^2(\beta)}, \quad (9)$$

with $\sigma_t^2(\beta) = \text{Var}[Y_t \mid \mathcal{F}_{t-1}]$. The solution of the score equation,

$$\mathbf{S}_N(\beta) = \nabla \log \text{PL}(\beta) = \mathbf{0} \quad (10)$$

is denoted by $\hat{\beta}$, and is referred to as the maximum partial likelihood estimator of β . The system of equations (10) is non-linear and is customarily solved by the Fisher scoring method, an iterative algorithm resembling the Newton-Raphson procedure. The *cumulative conditional information matrix*, $\mathbf{G}_N(\beta)$, defined by a sum of conditional covariance matrices,

$$\begin{aligned} \mathbf{G}_N(\beta) &= \sum_{t=1}^N \text{Cov} \left[\mathbf{Z}_{t-1} \frac{\partial \mu_t(Y_t - \mu_t(\beta))}{\partial \eta_t} \frac{1}{\sigma_t^2(\beta)} \mid \mathcal{F}_{t-1} \right] \\ &= \sum_{t=1}^N \mathbf{Z}_{t-1} \left(\frac{\partial \mu_t}{\partial \eta_t} \right)^2 \frac{1}{\sigma_t^2(\beta)} \mathbf{Z}_{t-1}' \\ &= \mathbf{Z}' \mathbf{W}(\beta) \mathbf{Z}. \end{aligned} \quad (11)$$

with

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_0' \\ \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_{N-1}' \end{bmatrix},$$

an $N \times p$ matrix, and $\mathbf{W}(\beta) = \text{diag}(w_1, \dots, w_N)$ where

$$w_t = \left(\frac{\partial \mu_t}{\partial \eta_t} \right)^2 \frac{1}{\sigma_t^2(\beta)}, \quad t = 1, \dots, N,$$

also an $N \times N$ diagonal matrix. The *unconditional information matrix* is simply

$$\text{Cov}(\mathbf{S}_N(\beta)) = \mathbf{F}_N(\beta) = \mathbf{E}[\mathbf{G}_N(\beta)].$$

Furthermore, let $\mathbf{H}_N(\boldsymbol{\beta})$ denote the matrix of second derivatives of the log-partial likelihood multiplied by -1:

$$\mathbf{H}_N(\boldsymbol{\beta}) \equiv -\nabla \nabla' l(\boldsymbol{\beta}).$$

Then

$$\mathbf{H}_N(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta}) - \mathbf{R}_N(\boldsymbol{\beta})$$

where

$$\mathbf{R}_N(\boldsymbol{\beta}) = \frac{1}{\alpha_t(\phi)} \sum_{t=1}^N \mathbf{Z}_{t-1} d_t(\boldsymbol{\beta}) \mathbf{Z}_{t-1}' (Y_t - \mu_t(\boldsymbol{\beta})),$$

and $d_t(\boldsymbol{\beta}) = [\partial^2 u(\eta_t) / \partial \eta_t^2]$. Here, we have used the notation $u(\cdot) \equiv (g \circ \mu(\cdot))^{-1} = \mu^{-1}(g^{-1}(\cdot))$ (see [7] for a full proof).

It turns out that canonical links entail certain simplifications. To see this, recall that when the link is canonical, $\eta_t = \theta_t$. Thus, the partial score (9) reduces to

$$\mathbf{S}_N(\boldsymbol{\beta}) = \frac{1}{\alpha_t(\phi)} \sum_{t=1}^N \mathbf{Z}_{t-1} (Y_t - \mu_t(\boldsymbol{\beta})),$$

and the conditional information matrix (11) is equal to

$$\mathbf{G}_N(\boldsymbol{\beta}) = \frac{1}{\alpha_t^2(\phi)} \sum_{t=1}^n \mathbf{Z}_{t-1} \sigma_t^2(\boldsymbol{\beta}) \mathbf{Z}_{t-1}',$$

Moreover, for canonical links it holds that $u(\eta_t) = \eta_t$ in which case $d_t = 0$. Therefore $\mathbf{R}_N(\boldsymbol{\beta})$ vanishes and we have the useful result

$$\mathbf{H}_N(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta}).$$

The discussion so far can be summarized in the following theorem, where the quantity $\mathbf{G}(\boldsymbol{\beta})$ is defined as the limiting conditional information matrix which is assumed to exist and be invertible at the true value of the parameter. The consequences of the theorem is that confidence intervals and testing are implemented by standard procedures.

Theorem 1 *Under Assumption A of [7], the maximum partial likelihood estimator is almost surely unique for all sufficiently large N , and it is consistent and asymptotically normal:*

$$\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}$$

in probability, and

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow \mathcal{N}_p(\mathbf{0}, \mathbf{G}^{-1}(\boldsymbol{\beta})),$$

in distribution, as $N \rightarrow \infty$, where $\mathcal{N}_p$ denotes the p -dimensional normal distribution.

4.1 Example

Consider a time series of counts where a random time dependent covariate is added in the model that incorporates moving averages components of the form (7). More specifically, suppose that

$$\log \mu_t(\boldsymbol{\beta}) = \beta_{01} + \beta_{02} X_t + \beta_1 e_{t-1} \quad (12)$$

for $t = 1, \dots, N$, and $X_t = \phi X_{t-1} + w_t$ where $|\phi| < 1$ and the sequence $\{w_t\}$, $t = 1, \dots, N$ forms a random sample from the normal distribution with mean 0 and

variance τ^2 . By setting $\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, \beta_1)'$ and $\mathbf{Z}_{t-1} = (1, X_t, e_{t-1})'$ it is obtained that $\log \mu_t = \eta_t = \mathbf{Z}_{t-1}'\boldsymbol{\beta}$, so that model (8) holds true. Table 1 reports limited simulation results for model (12) when $\boldsymbol{\beta} = (-1, -0.50, -0.50)$, $\phi = 0.5$ and $\tau^2 = 1$ for different values of the sample size N and the parameter λ . The reported numerical results are based on the average of 100 simulations and consist of 100 estimates $\hat{\beta}_i$, $i = 1, 2, \dots, 100$ from which sample means and variances are computed. The theoretical standard errors were calculated by inverting (11) and are reported in parentheses underneath the estimated standard errors. In most cases considered, the regression parameters are close to the true value and the simulation output points out to a close agreement between theory and empirical results, even for moderately small sample sizes, say $N = 100$. Figure 2 shows histograms of 100 maximum partial likelihood estimators under the same simulation scheme when $N = 100$. The superimposed curve is a standard kernel density estimate which reveals that the approximate distribution of $\hat{\boldsymbol{\beta}}$ is a symmetric distribution centered at the true values of the parameter vector.

Table 1: Maximum partial likelihood estimators, estimated and theoretical (in parentheses) standard errors for model (12) when $\boldsymbol{\beta} = (-1, -0.50, -0.50)$, $\phi = 0.5$ and $\tau^2 = 1$ using different N and λ . Results are based on 100 simulations.

N	λ	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_1$	SE		
100	0.50	-1.0096	-0.4961	-0.5524	0.1603 (0.1811)	0.1363 (0.1351)	0.2544 (0.2195)
200		-1.0118	-0.4939	-0.5227	0.0970 (0.1309)	0.0864 (0.0854)	0.1540 (0.1489)
300		-0.9933	-0.4854	-0.5053	0.0813 (0.1045)	0.0628 (0.0687)	0.1161 (0.1133)
100	1.00	-1.0147	-0.4955	-0.5385	0.2211 (0.1959)	0.1313 (0.1155)	0.2495 (0.1664)
200		-0.9997	-0.4892	-0.5118	0.1311 (0.1400)	0.0850 (0.0988)	0.1530 (0.1198)
300		-0.9993	-0.4933	-0.5135	0.0840 (0.1145)	0.0646 (0.0753)	0.1074 (0.0999)

5 Conclusions

The aim of this short communication is to show that two—of many other—ideas of Professor Cox changed our perception for time series analysis. On one hand it is the partial likelihood method that allows for sequential conditional inference and on the other hand it is the concept of observation driven models for prediction. Both of these ideas lay the foundations for time series analysis within the framework of generalized linear models. Hence, model fitting, deviance, predictions and residuals can be easily accessible to the data analyst. Although, the article specializes to time series following generalized linear models, both of the concepts can be employed to cover more general situations.

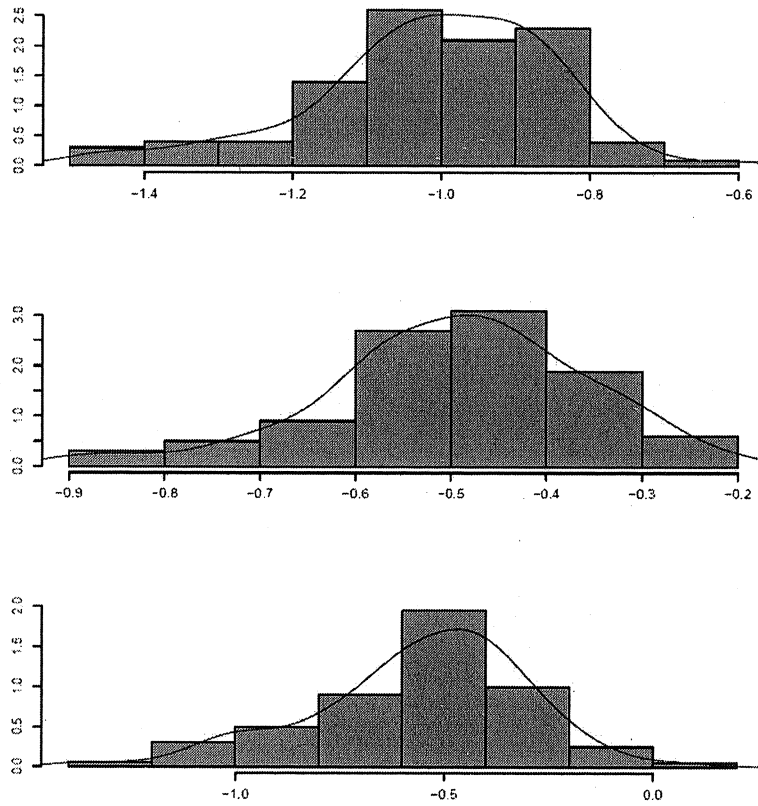


Figure 2. Histograms and density estimates from 100 maximum partial likelihood estimators when model (12) holds with $N = 100$, $\phi = 0.5$, $\tau^2 = 1$, $\lambda = 0.5$ and $\boldsymbol{\theta} = (-1, -0.50, -0.50)'$. From top to bottom: Histogram of $\hat{\beta}_{01}$, $\hat{\beta}_{02}$ and $\hat{\beta}_1$. Results are based on 100 simulations.

References

- [1] Benjamin, M. A., R. A. Rigby, and D. M. Stasinopoulos (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association* **98**, 214–223.
- [2] Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **74**, 187–220. with discussion.
- [3] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 69–76.
- [4] Cox, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics* **8**, 93–115.
- [5] Davis, R. A., Y. Wang, and W. T. M. Dunsmuir (1999). Modelling time series of count data. In S. Ghosh (Ed.), *Asymptotics, Nonparametric & Time Series*, pp. 63–114. New York: Marcel Dekker.
- [6] Fokianos, K. and B. Kedem (2004). Partial likelihood inference for time series following generalized linear models. *Journal of Time Series Analysis* **25**, 173–197.
- [7] Kedem, B. and K. Fokianos (2002). *Regression Models for Time Series Analysis*. Hoboken, NJ: Wiley.
- [8] Li, W. K. (1994). Time series models based on generalized linear models: some further results. *Biometrics* **50**, 506–511.
- [9] McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.
- [10] Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370–384.
- [11] Shephard, N. (1995). Generalized linear autoregression. Technical report, Nuffield College, Oxford University. available at www.nuff.ox.ac.uk/economics/papers/1996/w8/glar.ps.
- [12] Wong, W. H. (1986). Theory of partial likelihood. *Annals of Statistics* **14**, 88–123.

Some Applications of Bayesian Methods in Cancer Clinical Trials

Edmund A. Gehan

*Lombardi Comprehensive Cancer Center
Georgetown University Medical Center
Washington, DC 20057, USA
Email: gehane@georgetown.edu*

Abstract : This paper gives two applications of Bayesian methods in cancer clinical trials, the first to a 6MP versus placebo trial in acute leukemia also used by Cox [4] in his landmark paper on the proportional hazards model and the second to a trial of a new treatment (AMSA-OAP) in previously untreated patients with acute leukemia. The proposed strategy in the second application was to administer the new treatment to patients with poor predicted prognosis and a standard treatment to patients with good predicted prognosis, the predictions being based upon a logistic model fit to a historical control series of patients. Let θ_1 ($0 \leq \theta_1 \leq 1$) be the preference rate for the new versus the standard treatment based upon the responses between paired patients receiving the new treatment in the current clinical trial versus a comparable patient receiving the standard treatment in the historical control series. Let θ_2 ($0 \leq \theta_2 \leq 1$) be the preference rate for the standard treatment in the current trial versus paired patients receiving the standard treatment in the historical control series. Then calculating the posterior probabilities ($\theta_i > 1/2 \mid \text{data}$) and ($\theta_i = 1/2 \mid \text{data}$) for $i = 1, 2$ provide a quantitative basis for deciding whether the new treatment is better than the standard in patients with poor predicted prognosis and whether the standard treatment has the same response rate in the current trial as in the historical control series in patients with good predicted prognosis. The proposed strategy and the calculation of the relevant posterior probabilities provide an ethical solution to the problem of evaluating a new treatment in a disease when effective treatments already exist.

Key words: Bayesian methods, cancer clinical trials.

Personal Remarks

It was a distinct pleasure to participate in the conference honouring the 80th birthday of Sir David Cox. I first met David when he was a visitor to the Department of

Biostatistics at the University of North Carolina in 1955-56. I was assigned by the then Head of the Department, Bernard Greenberg, to help David adjust to life in America. I helped by assisting in his purchase of a used car and teaching him to drive. I was surprised, about five years later when visiting London, that he was still driving the same car, a light blue Chevrolet, with the steering wheel on the left-hand side. As in other fields, he had progressed much beyond his teacher in his driving ability. When driving with him on a curved road outside London, he showed me how to pass an auto on a blind curve. He achieved this by drifting his car further to the left, so he had a longer line of vision beyond the left-hand side of the car in front. Seeing no on-coming traffic, he confidently passed the auto on the blind curve, but his old teacher was rather shaken in the process.

The members of the Biostatistics Department and I also introduced David to April Fool's Day as celebrated in the U.S. We knew about his love of fresh air whatever the temperature, so when he arrived at his office on April 1 in the World War II shack that was then the home for the Department, he found that the windows were taped shut. There was a notice taped to a window:

"A randomized study is in progress to investigate the effect of fresh air on research output and you have been assigned to the control group to receive NO fresh air." So I can vouch that he is a good driver of automobiles (not yet having had an accident) and also a good sport, in addition to being an outstanding statistician.

1 Introduction

When I worked in Professor Cox's Department of Statistics at Birkbeck College in the early 1960's, I discussed a clinical trial of 6-mercaptopurine (6MP) versus placebo in acute leukemia that was designed to test the hypothesis that 6MP was superior to placebo in increasing the length of complete remission. This trial was used as an example in a paper that I worked on generalizing the Wilcoxon test [6], and later Professor Cox used the same example in his landmark 1972 paper on survival analysis [4]. I have since been gratified to hear him credit this trial as one of the influences leading to the development of the proportional hazards model.

In this paper, I propose to reanalyze the 6MP versus placebo trial from the Bayesian viewpoint. Second, I propose to utilize the same Bayesian approach as a strategy for evaluating a new treatment in previously untreated patients with acute leukemia. The problem of evaluating a new treatment in a disease when effective treatments already exist is a difficult one, since ethics suggest administering the known effective treatments first and any new treatment later in the course of disease after the known treatments have failed. Scientifically however, it is believed that the best opportunity for a new treatment is in the newly diagnosed patient, rather than later in the course of disease when the patient may have developed a resistance to almost any therapy. Hence, to evaluate a new treatment in previously untreated patients, one must balance

the ethical requirement of giving known effective treatments first against the scientific belief that a new treatment is likely to be most effective early in the course of disease.

2 Clinical Trial of 6-Mercaptopurine (6MP) versus Placebo on the Duration of Remissions in Acute Leukemia

This trial was reported in [5] and was designed as follows. Patients received steroid treatment until achieving a complete or partial bone marrow remission and were then randomized within pairs of patients, separately by institution and remission status, to maintenance of the remission with 6MP or placebo. In the planning stage, there were strong vocal arguments that almost prevented the trial from being conducted. Those believing that 6MP was effective thought that it was unethical to randomize patients to placebo, while those arguing for the placebo randomization said that 6MP had never been tested for its ability to maintain remissions. In the discussion, two points led to the trial being approved. First, the trial was to be conducted sequentially according to one of the [1] plans, so the trial would be stopped early, if 6MP was substantially superior to placebo. Second, patients relapsing on placebo would then receive 6MP treatment. Hence, patients would receive 6MP either at the beginning of their remission or, if relapsing on placebo, at the time of relapse. Table 1 gives the actual lengths of remission (in weeks) for the patients, the drug assignment, and the preferences for 6MP or placebo, the preferences being decided by the treatment that led to the longer remission. The data show a predominance of preferences for 6MP and, according to the wedge sequential plan of [1], a decision to stop the trial could be reached after 18 pairs of patients, at which time there were 15 preferences for 6MP and three for placebo. The Armitage restricted sequential procedure was designed to be sensitive to a proportion of 6MP or placebo preferences of 0.75 and, if an upper or lower limit was passed, statistical significance could be reported at $P = 0.05$. There were three additional pairs of patients (19, 20, 21) available at the time of stopping the trial, since the trial was analyzed at approximate four month intervals. A simplified version of the data in Table 1 was reported in [4], assuming a randomized clinical trial of 42 patients, 21 patients randomized to each treatment without consideration of the pairing or sequential nature of the trial.

3 Bayesian Analysis of the Acute Leukemia Trial

Let θ be the preference rate for 6MP among pairs of patients and $(1 - \theta)$ the preference rate for placebo, where $(0 \leq \theta \leq 1)$,

$\pi(\theta)$ = prior distribution for θ based upon data or beliefs about θ

and

$$\begin{aligned} p(\mathbf{x} | \theta) &= \text{probability distribution of data } \mathbf{x}, \text{ for given } \theta \\ &= \frac{n!}{r!} (n-r)! \theta^r (1-\theta)^{n-r} \quad (0 \leq \theta \leq 1) \end{aligned}$$

when there are n patients, and r preferences for 6MP.

By Bayes theorem,

$$\pi(\theta | \mathbf{x}) = \frac{\pi(\theta) p(\mathbf{x} | \theta)}{p(\mathbf{x})}$$

and

$$p(\mathbf{x}) = \int_0^1 \pi(\theta) p(\mathbf{x} | \theta) d\theta$$

where $\pi(\theta)$ is the prior distribution of θ and $p(\mathbf{x} | \theta)$ is the likelihood of the data given θ .

The probability distribution of the data ($\mathbf{x}$) is binomial and different possibilities could be considered as choices for $\pi(\theta)$. First, consistent with the approach of Cornfield [3] I will assume a non-zero prior probability p that $\theta = 1/2$, so that 6MP has the same prior response rate as placebo and also that the remaining probability $(1-p)$ is distributed uniformly over $\theta \neq 1/2$. Cornfield [3] was the first to consider this approach as a way of avoiding sampling to a foregone conclusion. Second, I consider a uniform prior distribution on θ in $(0, 1)$ and am interested in the posterior probability that 6MP is equal in response rate to placebo or the better treatment. These considerations lead to:

If there is a non-zero prior probability p that $\theta = 1/2$, $1-p$ is distributed uniformly over $\theta \neq 1/2$ and there are r_{6MP} , r_p preferences for 6MP, placebo in a sample of n patients, then

$$\text{Prob}(\theta = 1/2 | r_{6MP}, r_p) = \left[1 + \left(\frac{1-p}{p} \right) 2^{r_{6MP}+r_p} \int_0^1 \theta^{r_{6MP}} (1-\theta)^{r_p} d\theta \right]^{-1} \quad (1)$$

gives the posterior probability that the treatments are equally preferred.

Assuming a uniform prior for θ in $(0, 1)$, then

$$\text{Prob}(\theta > 1/2 | r_{6MP}, r_p) = \frac{(r_{6MP} + r_p + 1)!}{r_{6MP}! r_p!} \int_{1/2}^1 \theta^{r_{6MP}} (1-\theta)^{r_p} d\theta \quad (2)$$

gives the posterior probability that 6MP is preferred.

This Bayesian approach has been considered by Berry [2], and Table 2 gives the relevant posterior probabilities, the nominal two-sided P value (assuming a fixed sample size study), and the sequential two-sided significance level [1]. Since there was a prior belief that 6MP would be the better treatment, the posterior probability that $(\theta > 1/2 | \text{data})$ is the most relevant column. At patient pair 10, the posterior probability that 6MP is better reached 0.967, a time at which the trial could have been stopped. At the 18th pair of patients when the trial could actually have been stopped according to the Armitage plan, the posterior probability that $(\theta > 1/2 | \text{data})$ was 0.998 and the posterior probability that $(\theta = 1/2 | \text{data})$ was only 0.06, very similar to the sequential two-sided significance level of 0.05. In the next section, I will consider

utilizing this same Bayesian approach in a clinical trial evaluating a new treatment in acute leukemia.

4 Strategy for Evaluation of a New Treatment in Untreated Patients with Acute Leukemia

A clinical trial of a new treatment, AMSA-OAP, in previously untreated patients with acute leukemia was reported by Keating, Gehan *et. al.* [7]. This clinical trial involved an innovative strategy for the assignment of patients to treatment based upon predicted probabilities of complete remission (p_r values) derived using six prognostic factors in a logistic regression model. The six prognostic factors in the model were: age group, antecedent hematologic disorder (yes or no), body temperature, BUN status, hemoglobin value, and liver size.

A historical control group of 325 patients was available that received adriamycin-OAP (AD-OAP) which was considered the standard treatment from 1973 – 1977. The complete remission rate in these patients was 61%.

The strategy for the conduct of the trial can be summarized as follows: administer the new treatment (AMSA-OAP) to patients with poor predicted prognosis and the standard treatment (AD-OAP) to patients with good predicted prognosis based on the logistic model fit to the historical control data on the standard treatment. In the initial stage of the trial, patients with poor predicted prognosis were defined as those with $p_r < 0.4$, and the responses in these patients were compared with paired patients with the same p_r values (± 0.02) who received AD-OAP therapy from 1973 – 1977. Patients with good predicted prognosis were those with $p_r \geq 0.4$ and these patients received the standard treatment (AD-OAP). Responses were compared with paired patients with the same p_r values (± 0.02) who received AD-OAP therapy in the historical control series. In this application, θ_1 is the preference rate for AMSA-OAP among pairs of patients with poor predicted prognosis and θ_2 is the preference rate for AD-OAP among pairs of patients with good predicted prognosis. For both groups of patients ($i = 1, 2$), posterior probability calculations could be made of $\text{Prob}(\theta_i = 1/2 \mid \text{data})$ and $\text{Prob}(\theta_i > 1/2 \mid \text{data})$ as in (1) and (2). For patients with poor predicted prognosis, high values of $\text{Prob}(\theta_1 > 1/2 \mid \text{data})$ would be evidence that AMSA-OAP was superior in response rate to AD-OAP in the historical control series. For patients with good predicted prognosis, high values of $\text{Prob}(\theta_2 = 1/2 \mid \text{data})$ would be evidence that the response rate to AD-OAP in the current trial was the same as that in the historical control series.

The trial was ultimately conducted in three stages with an expanding criterion for administering AMSA-OAP in each stage. Because the results for AMSA-OAP were favorable in the initial stages, the criterion for administering AMSA-OAP increased with each stage: Stage 1: $p_r < 0.40$; Stage 2: $p_r < 0.6$; Stage 3: $p_r \leq 1.0$. The actual trial was conducted using a frequentist criterion for determining the change

in strategy by stage, but the results are here analyzed from the Bayesian viewpoint. Table 3 gives the cumulative results of the trial by Stage so that the number of patients (50) reported in Stage 3 for the AMSA-OAP subgroup $p_r < 0.40$ includes all patients entered through the three stages of the trial and the associated statistics refer to the cumulated results. At the end of Stage 1, the posterior probability ($\theta_1 > 1/2$ | data) in patients with $p_r < 0.40$ reached 0.93, a reasonable posterior probability for deciding to increase the criterion for administering AMSA-OAP. For patients receiving AD-OAP in Stage 1, i.e., those with $p_r \geq 0.40$, the posterior probability ($\theta_2 = 1/2$ | data) achieved 0.81, reasonably strong evidence that the complete remission rate for AD-OAP was the same in the current trial as in the historical control series; the complete remission rates in Stage 1 were 70%, 71%, respectively.

Results for AMSA-OAP continued to be favorable in Stage 2 with a Prob ($\theta_1 > 1/2$ | data) of 0.95 in patients with $p_r < 0.40$ and 0.83 in patients with $0.40 \leq p_r \leq 0.60$. Hence, the criterion for administering AMSA-OAP was increased to ($p_r \leq 1.0$) in Stage 3. At the end of Stage 3, there was strong evidence that AMSA-OAP was superior in response rate to AD-OAP in the 50 patients with ($p_r < 0.40$) since Prob ($\theta_1 > 1/2$ | data) was 0.95. In the 41 patients with ($0.40 \leq p_r \leq 1.0$) at the end of Stage 3, there was moderate evidence of the superiority of AMSA-OAP since Prob ($\theta_1 > 1/2$ | data) was 0.88. Though the Bayesian analysis suggested a strong preference for AMSA-OAP in patients with ($p_r < 0.40$) and a moderate preference in patients with ($0.40 \leq p_r \leq 1.0$), the standard P values were only 0.15 and 0.36 in these subgroups at the end of the third stage, both P values indicating a non-significant difference between the treatments.

Among patients receiving AD-OAP in the current study, ultimately there were 108 patients at the end of Stage 2 who received AD-OAP and preferences could be recorded in 45 pairs of patients. The complete remission rate at the end of Stage 2 was 72% for AD-OAP compared with 71% in the patients in the historical control group (Table 3). The posterior probability that $\theta_2 = 1/2$ was 0.84, quite strong evidence of the equivalence in response rates between AD-OAP in the current series of patients and AD-OAP in the historical control series.

5 Discussion

The proposed strategy for evaluation of new treatments in previously untreated patients in cancer is relevant when the primary endpoint is response rate, some historical data are available for a standard treatment, and it is reasonable to fit a logistic model accounting for relevant prognostic factors to obtain predicted probabilities of response on the standard treatment.

The proposed strategy for the trial is to conduct a historical control study and administer the new treatment to patients with poor predicted prognosis and the standard treatment to patients with good predicted prognosis based upon the logistic model. This strategy achieves a balance between the ethical requirement of administering a

standard treatment to patients likely to benefit from it, and the scientific belief that a new treatment is best evaluated in previously untreated patients. The proposed strategy is ethical, since it involves administering the new treatment only to patients unlikely to benefit from the standard treatment. One should also be aware that before the conduct of a comparative trial, a new treatment would previously have been studied in a Phase I trial to determine a maximum tolerated dose and a Phase II trial to determine that the therapy had a sufficiently high response rate to be advanced to a Phase III comparative trial.

Another requirement for this strategy is the availability of a sufficiently large series of historical control patients who had received the standard treatment. The conditions for patients receiving the new treatment in the current trial should be sufficiently similar to those in the historical control series so that the predictions of prognosis from the logistic model are valid. Ideally, the historical control series should be recent, involve the same types of patients, the same criteria for achieving response, and have comparable eligibility requirements for being included in the trial. Utilizing a Bayesian approach permits the calculation of relevant posterior probabilities for determining when to adjust the strategy for administering the new treatment.

The Bayesian approach can achieve two objectives in the proposed strategy. First, when the new treatment is administered to patients with poor predicted prognosis, calculating the posterior probability ($\theta_1 > 1/2 \mid \text{data}$) and the posterior probability ($\theta_1 = 1/2 \mid \text{data}$) provides a quantitative basis for deciding whether the new treatment is equal to or better than the standard or not. Secondly, in the group of patients receiving the standard treatment, the Bayesian posterior probability for ($\theta_2 = 1/2 \mid \text{data}$) provides a quantitative basis for deciding whether the current series of patients is similar in response rate to the historical control series. This data is important for deciding whether the historical control series provides a valid basis for applying the proposed strategy to the evaluation of new treatments in cancer.

Acknowledgments: I thank Olga Cobbs for assistance in the preparation of this manuscript.

References

- 1] Armitage P. (1957). Restricted sequential procedures. *Biometrika* 44:9-26.
- 2] Berry DA (1987). Interim analysis in clinical trials: the role of the likelihood principle. *The American Statistician*, 41, 2, 117-122.
- 3] Cornfield J. (1966). A Bayesian test of some classical hypotheses-with applications to sequential clinical trials. *J. of the American Statistical Association* 61, 315, 577-594.
- 4] Cox DR (1972). Regression models and life tables. *J. Royal Stat. Soc. B* 34: 187-220.
- 5] Freireich EJ, Gehan EA, et al (1963). The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia. *Blood* 21:699-716.

- [6] Gehan EA (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 52:203-222.
- [7] Keating MJ, Gehan EA, et al (1987). A strategy for evaluation of new treatments in untreated patients: application to a clinical trial of AMSA for acute leukemia. *Journal of Clinical Oncology* 5, 5, 710-721.

Pair	Remission Status	Drug	Length of Remission (wks.)	Preference
1	partial	placebo	1	6MP
	partial	6MP	10	
2	complete	placebo	22	placebo
	complete	6MP	7	
3	complete	placebo	3	6MP
	complete	6MP	32+	
4	complete	placebo	12	6MP
	complete	6MP	23	
5	partial	placebo	8	6MP
	partial	6MP	22	
6	partial	placebo	17	placebo
	partial	6MP	6	
7	complete	placebo	2	6MP
	complete	6MP	16	
8	complete	placebo	11	6MP
	complete	6MP	34+	
9	partial	placebo	8	6MP
	partial	6MP	32+	
10	complete	placebo	12	6MP
	complete	6MP	25+	
11	complete	placebo	2	6MP
	complete	6MP	11+	
12	partial	placebo	5	6MP
	partial	6MP	20+	
13	complete	placebo	4	6MP
	complete	6MP	19+	
14	complete	placebo	15	placebo
	complete	6MP	6	
15	complete	placebo	8	6MP
	complete	6MP	17+	
16	partial	placebo	23	6MP
	partial	6MP	35+	
17	partial	placebo	5	6MP
	partial	6MP	6	
18	complete	placebo	11	6MP
	complete	6MP	13	
19	complete	placebo	4	6MP
	complete	6MP	9+	
20	complete	placebo	1	6MP
	complete	6MP	6+	
21	complete	placebo	8	6MP
	complete	6MP	10+	

Table 1: Preferences for 6MP and Placebo in Sequential Phase of Study (Phase II).

Patient Pair	Preferred Treatment	r_{6MP}	r_p	Sequential 2-sided sig. level	Nominal 2-sided P value	Posterior Prob. $\theta = 1/2$ ¹	Posterior prob. 6MP better
1	6MP	1	0	-	1.0	0.50	0.75
2	placebo	1	1	-	1.0	0.60	0.50
3	6MP	2	1	-	1.0	0.60	0.69
4	6MP	3	1	-	0.62	0.56	0.81
5	6MP	4	1	-	0.38	0.48	0.89
6	placebo	4	2	-	0.69	0.62	0.77
7	6MP	5	2	-	0.45	0.58	0.86
8	6MP	6	2	-	0.29	0.50	0.91
9	6MP	7	2	-	0.18	0.41	0.945
10	6MP	8	2	-	0.11	0.33	0.967
11	6MP	9	2	-	0.06	0.24	0.981
12	6MP	10	2	-	0.04	0.17	0.989
13	6MP	11	2	-	0.02	0.12	0.994
14	placebo	11	3	-	0.06	0.25	0.982
15	6MP	12	3	-	0.04	0.18	0.989
16	6MP	13	3	-	0.02	0.13	0.994
17	6MP	14	3	-	0.01	0.08	0.997
18	6MP	15	3	0.05	0.008	0.06	0.998
19	6MP	16	3	-	0.004	0.03	0.999
20	6MP	17	3	-	0.003	0.02	0.999
21	6MP	18	3	-	0.002	0.01	0.999

Table 2: Analysis of Sequential Trial of 6MP vs Placebo in Acute Leukemia².

Treatment Group		Stage		
		1	2	3
AMSA-OAP ($p_r < 0.40$)	No. Patients	21	38	50
	No. Preferences	12	19	24
	CR% (AMSA-OAP)	43%	39%	38%
	CR% (AD-OAP) ³	14%	21%	22%
	Prob ($\theta_1 > 1/2$ data)	0.93	0.95	0.95
	Prob ($\theta_1 = 1/2$ data) ¹	0.44	0.51	0.52
	P	0.15	0.17	0.15
AMSA-OAP ($0.40 \leq p_r < 0.60$), Stage 2 ($0.40 \leq p_r \leq 1.0$), Stage 3	No. Patients	-	22	41
	No. Preferences	-	11	19
	CR% (AMSA-OAP)	-	50%	50%
	CR% (AD-OAP) ³	-	36%	44%
	Prob ($\theta_1 > 1/2$ data)	-	0.83	0.88
	Prob ($\theta_1 = 1/2$ data) ¹	-	0.66	0.66
	P	-	0.55	0.36
AD-OAP ($p_r \geq 0.40$), Stage 1 ($p_r \geq 0.60$), Stage 2	No. Patients	63	108	-
	No. preferences	28	45	-
	CR% (AD-OAP)	72%	73%	-
	CR% (AD-OAP) ³	71%	71%	-
	Prob ($\theta_2 > 1/2$ data)	0.29	0.62	-
	Prob ($\theta_2 = 1/2$ data) ¹	0.81	0.84	-
	P	0.85	0.88	-

Table 3: Results of AMSA-OAP vs AD-OAP (Historical Control) Trial in Three Stages.

¹ Assumes $p = 1/2$ in equation (1).² Modified from Berry [2].³ Complete remission rate in paired patients on AD-OAP in historical control group.

On Potential Risk Reversal Due to Heterogeneity

David Oakes

*Department of Biostatistics and Computational Biology
University of Rochester
Rochester, NY 14642, USA
Email: oakes@bst.rochester.edu*

Abstract : Prompted by unexpected results from a study of the influence of coronary disease-related genotypes on recurrent cardiac events among patients who had already experienced a myocardial infarction, we consider the potential for risk reversal due selection in a heterogeneous population. We discuss two possible mechanisms, frailty selection and frailty enrichment. Under a simple frailty enrichment model in which the total risk for an event is the sum of an observed and an unobserved component, we show that among those who have experienced a first event, those with the higher observed risk may be at lower risk of a second event, even if, at an individual level, the risks for the second event are the same as those for the first event.

Key words: frailty, length-biased sampling, multiple events, Simpson's paradox.

1 Introduction

In a recent study, Moss *et al.* (2005) hypothesized that genotypes found previously to be associated with increased risk of a coronary event would be associated also with increased risk for a subsequent event among subjects who had already experienced a first event. Contrary to this hypothesis, it was found that the risk of a subsequent event was significantly decreased among the subjects possessing these genotypes. While a variety of explanations of this surprising phenomenon can be proposed, it prompted us to explore mathematical models that might exhibit such risk reversal. Two such mechanisms may be acting here - in order to experience a second event one must (a) have experienced a first event (we call this *frailty enrichment*) and (b) have survived it (we call this *frailty selection*). Both models postulate the existence of unmeasured risk factors resulting in heterogeneity of risk in the population beyond that explicable

by measured explanatory variables. Of the two mechanisms, frailty selection is the more familiar and we consider it first.

2 Frailty selection

This class of models, sometimes colloquially termed as “the survival of the fittest” has been well described by Vaupel and Yashin (1985). A specific example often quoted is that death rates for smokers may be lower than those for nonsmokers at certain ages, because the surviving smokers will have been subject to more rigorous selection than surviving nonsmokers.

To illustrate the phenomenon mathematically we first consider the following simple model. Suppose that a population consists of equal portions of susceptible individuals, whose survival times are exponentially distributed with scale parameter $\rho = 2$ and immune individuals, whose survival times are exponentially distributed with parameter $\rho = 1$. Suppose also that the effect of exposure to a certain risk factor on “susceptible” individuals is to increase $\rho = 2$ to $\rho' = 3$, but that exposure to this risk factor has no effect on “immune” individuals.

The overall survivor function for the “unexposed” population, namely

$$S_U(t) = \frac{1}{2}(e^{-2t} + e^{-t}),$$

exceeds that for the “exposed” population,

$$S_E(t) = \frac{1}{2}(e^{-3t} + e^{-t}),$$

for all t . The same is not true for the corresponding hazard functions,

$$h_U(t) = -\frac{S'_U(t)}{S_U(t)} = \frac{1 + 2e^{-t}}{1 + e^{-t}}$$

and

$$h_E(t) = -\frac{S'_E(t)}{S_E(t)} = \frac{1 + 3e^{-2t}}{1 + e^{-2t}},$$

It is easily seen that $h_E(t) > h_U(t)$ if $t < -\log(\sqrt{2} - 1) = 1.51 = t_0$ say, but that for $t > t_0$, $h_E(t) < h_U(t)$. If we just consider the marginal hazard functions among the overall surviving population, without consideration of the confounding effects of susceptibility, the harmful exposure would appear beneficial for $t > t_0$.

A related phenomenon is that the optimal weight function for the weighted logrank statistic for testing the hypothesis $\beta = 0$ in the multiplicative frailty model

$$h(t, \beta) = W e^{\beta z} h_0(t), \quad (1)$$

may be negative for certain values of t . Here β is the coefficient of the observed covariate z , W is an unobserved random variable whose distribution has Laplace transform $p(s) = E(e^{-sW})$ and $h_0(t)$ is an unspecified baseline hazard function. See e.g. Oakes and Jeong (1998, p. 216).

It is, of course, well known that the hazard functions of two distributions may cross without the corresponding survivor functions crossing, (see e.g. Cox and Oakes, 1984, p. 79) and the frailty selection effect may be thought of in these terms.

We considered, but ultimately rejected, frailty selection as a possible explanation for the findings of Moss et al., because numerical calculations showed that mortality from first heart attacks was not sufficient to explain the risk reversal we had observed.

3 Frailty enrichment - absolute risk model

Unlike the classic frailty selection model (1), which postulates that the observed co-variate and unobserved frailty W combine multiplicatively in the hazard function, the frailty enrichment approach is best exemplified by an additive combination of the observable and unobservable risks. We first consider an absolute risks model in which the observable binary risk factor z conveys a positive risk θ which combines additively with the unobservable risk u to give total risks of u and $u + \theta$ for an individuals with $z = 0$ and $z = 1$ respectively.

Suppose that u varies in the population according to a probability density function $f(u)$ with mean and variance

$$\mu = \int_0^1 u f(u) du, \quad \sigma^2 = \int_0^1 u^2 f(u) du - \mu^2.$$

Then the average (mean) risks among individuals with $z = 0$ and $z = 1$ are respectively μ and $\mu + \theta$ respectively.

Now suppose that individuals who have experienced a first event then become at risk for experiencing a second event. For any individual the risk of the second event is the same as that of the first event, i.e. each individual keeps the same value of u for the second event as the first. The distribution of the unobserved risks among subjects who have experienced a first event will differ from $f(u)$: it can be calculated from Bayes' theorem as

$$f_0(u) = \frac{f(u) \text{pr(First Event } | u)}{\int f(u) \text{pr(First Event } | u) du} = \frac{u f(u)}{\mu},$$

or subjects with $z = 0$, and similarly,

$$f_1(u) = \frac{u + \theta}{\mu + \theta} f(u),$$

or subjects with $z = 1$. The mean risk for the first group is

$$\mu_0 = \int u f_0(u) du = \int \frac{u^2 f(u) du}{\mu} = \mu + \frac{\sigma^2}{\mu},$$

and for the second group is

$$\mu_1 = \theta + \int u f_1(u) du = \theta + \frac{\int (u^2 + \theta u) f(u) du}{\mu + \theta} = \mu + \theta + \frac{\sigma^2}{\mu + \theta}.$$

These results are a minor variant of the well-known length-biased sampling phenomenon (Cox, 1967).

The difference in risks between individuals with $z = 1$ and $z = 0$ is

$$\mu_1 - \mu_0 = \theta - \frac{\sigma^2 \theta}{\mu(\mu + \theta)}.$$

This is always less than θ and may be negative if σ^2 is large. The value of the derivative of μ_1 in θ at $\theta = 0$ equals $1 - \sigma^2/\mu^2$ which is negative for distributions with coefficient of variation exceeding unity. This suggests that risk reversal for small values of θ may be quite a common phenomenon.

For a simple specific example of this risk reversal, suppose that u is dichotomous with 90% of subjects having $u = 0.01$ and 10% having $u = 0.20$, and that $\theta = 0.01$. Then among subjects with $z = 0$ the proportion of high-risk ($u = 0.2$) among those who have experienced a first event is

$$\frac{0.2 \times 10\%}{(0.20 \times 10\%) + (0.01 \times 90\%)} = 69.0\%.$$

Among subjects with $z = 1$ and who have experienced a first event the corresponding proportion is

$$\frac{0.21 \times 10\%}{(0.21 \times 10\%) + (0.02 \times 90\%)} = 53.8\%.$$

The risks of a second event among these subjects are

$$(69.0\% \times 0.2) + (31.0\% \times 0.01) = 0.141$$

among subjects with $z = 0$ and

$$(53.8\% \times 0.21) + (46.2\% \times 0.02) = 0.122$$

among subjects with $z = 1$.

A few comments are in order:

(1). This discrete risks model is subject to the constraint $u + \theta < 1$. The time-dependent intensity-based model in the next section avoids this constraint.

(2). The assumption that the risk of the second event for any individual is the same as the risk for the first event is made for convenience only: - the results would in essence be unchanged if all risks pertaining to second events were set to a constant multiple of the risks for first events.

(3). As noted above, risk reversal does not occur if the risk from z combines multiplicatively with the background risks (giving total risks of u for those with $z = 0$ and θu for those with $z = 1$).

4 Frailty enrichment - Intensity model

We now consider an additive model for the intensity of events. Let us suppose that an individual with unobserved risk - "frailty" - w experiences events in a homogeneous Poisson process of rate $\lambda + w$. Here λ is a function of the observable risk factor z . Then, if $N(t)$ denotes the number of events experienced in time t , we have

$$\text{pr}\{N(t) = 0\} = \int e^{-(\lambda+w)t} f(w) dw = e^{-\lambda t} f^*(t),$$

and

$$\text{pr}\{N(t) = 1\} = \int (\lambda + w)te^{-(\lambda+w)t}f(w)dw = \lambda te^{-\lambda t}f^*(t) - te^{-\lambda t}f^{*'}(t),$$

where $f(w)$, $f^*(t)$ and $f^{*'}(t)$ denote respectively the density of W , the Laplace transform of this density, and the derivative of the transform.

The conditional densities of W given $N(t) = 0$ and $N(t) = 1$ are respectively

$$f_0(w) = \frac{f(w)e^{-wt}}{f^*(t)},$$

and

$$f_1(w) = \frac{f(w)(\lambda + w)e^{-wt}}{\lambda f^*(t) - t f^{*'}(t)} = \frac{(\lambda + w)f_0(w)}{\int (\lambda + w)f_0(w)dw}.$$

The conditional intensity of a second event among subjects who have experienced a first event is

$$E(W|N(t) = 1) = \mu + \lambda + \frac{\sigma^2}{\mu + \lambda},$$

where now $\mu = \int w f_0(w)dw$ and $\sigma^2 = \int w^2 f_0(w)dw - \mu^2$ relate to the conditional density $f_0(w)$ rather than the original density $f(w)$. The algebra here, and the conclusions, are exactly analogous to that in the absolute risks model of Section 3.

As a simple example, suppose that

$$f(w) = \frac{\rho^\kappa w^{\kappa-1} e^{-\rho w}}{\Gamma(\kappa)},$$

the gamma density $\Gamma(\kappa, \rho)$ with index κ and scale parameter ρ . Then

$$f_0(t) \sim \Gamma(\kappa, \rho + t),$$

so that

$$\mu = \frac{\kappa}{\rho + t}, \quad \sigma^2 = \frac{\kappa}{(\rho + t)^2},$$

and the relevant coefficient of variation is κ^{-1} , irrespective of t . Hence risk reversal will always occur for small values of λ if $\kappa < 1$.

As in the absolute risks model, all risks for second events may be multiplied by the same constant compared to those for first events, without affecting the algebra.

Possible variants of the calculation are to examine the conditional intensity given $N(t) \geq 0$ instead of given $N(t) = 1$ or to condition just on the occurrence of an event at or very close to a specific time t . The second formulation leads back to exactly the same algebra as above, but in terms of the moments of $f(w)$ instead of those of $f_0(w)$. The first formulation leads to slightly different algebra, but the qualitative phenomenon of risk reversal may still occur.

5 Discussion

Whether the risk reversal observed in Moss *et al.* (2005) is actually due to the selection or enrichment effects or has some more mundane explanation remains an open question, which will require further investigation and replication of the numerical re-

sults on independent data sets. Irrespective of the final verdict as far as this study is concerned, risk reversal due to unobserved heterogeneity should be borne in mind as a possible explanation for differing risk profiles among subjects experiencing first and subsequent events.

Acknowledgements

I thank Arthur Moss, Daniel Ryan and the other members of the “Thrombo” team for much useful discussion.

References

- [1] Cox, D.R. (1967). Some sampling problems in technology, in *New Directions in Survey Sampling* (eds. N.L. Johnson and H. Smith). New York, Wiley. pp.506–527.
- [2] Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. London, Chapman and Hall.
- [3] Moss, A. J., Ryan, D., Oakes, D., Goldstein, R.E., Greenberg, H., Bodenheimer, M.M., Brown, M.W., Case, R.B., Dwyer, E.M., Eberly, S.W., Francis, C.W., Gillespie, J.A., Krone, R.J., Lichstein E., MacCluer, J.W., Marcus, F.I., McCarthy, J., Sparks, C.E. and Zareba, W. (2005), Atherosclerotic risk genotypes and recurrent coronary events after myocardial infarction. *American Journal of Cardiology*, **96**, 177–182.
- [4] Oakes, D. and Jeong, J.-H. (1998). Frailty models and rank tests. *Lifetime Data Analysis*, **4**, 209–228.
- [5] Vaupel, J.W. and Yashin, I.A. (1985). Heterogeneity’s Ruses: some surprising effects of selection on population dynamics. *The American Statistician*, **39**, 176–185.

Large-Deviation Efficiency of First and Second Order

Masafumi Akahira

*Institute of Mathematics
University of Tsukuba
Ibaraki 305-8571, Japan
Email: akahira@math.tsukuba.ac.jp*

Abstract : The first and second order large-deviation efficiency is discussed from a different viewpoint from the Bahadur efficiency with the lower bound for the tail probability of estimators depending on the Kullback-Leibler (K-L) information. The first and second lower bounds for the tail probability of an asymptotically median unbiased estimator are directly derived without the K-L information. Examples in the normal and two-sided exponential cases are also given.

Key words: Large-deviation efficiency, testing hypothesis, most powerful test, Cornish-Fisher expansion, saddlepoint approximation.

1 Introduction

In statistical inference, the higher order asymptotic efficiency of estimators is usually discussed in the sense of maximizing the asymptotic concentration probability of estimators around the true parameter, using the Edgeworth expansion (see, e.g. Akahira and Takeuchi (1981), Pfanzagl and Wefelmeyer (1985), Ghosh (1994), and Barndorff-Nielsen and Cox (1994)). And, for example, it is known that, under suitable regularity conditions, a bias-adjusted maximum likelihood estimator has the third order asymptotic efficiency in some class of estimators.

On the other hand, the Bahadur efficiency of estimators is well known as a concept from the viewpoint of large-deviation with the lower bound for the tail probability (see Bahadur (1971)). The bound for the tail probability of a consistent estimator is given by the form of K-L information, and the Bahadur type second order efficiency based on the asymptotic expansion of the amount of K-L information is discussed by Fu (1982) (see also Akahira (1995)).

In this paper, we discuss the first and second order large-deviation efficiency from a different viewpoint from the Bahadur efficiency. The first and second order lower bounds for the tail probability of an asymptotically median unbiased estimator are directly derived without the K-L information. Examples in the normal and two-sided exponential cases are also given.

2 Definitions

Suppose that $X_1, X_2, \dots, X_n, \dots$ is a sequence of independent and identically distributed (i.i.d.) random variables with a probability density function (p.d.f.) $f(x, \theta)$ with respect to a σ -finite measure μ , where $\theta \in \Theta$ and Θ is an open interval in $\mathbb{R}^1$. Put $\mathbf{X} := (X_1, \dots, X_n)$. If an estimator $\hat{\theta}_n := \hat{\theta}_n(\mathbf{X})$ of θ satisfies

$$P_{\theta,n}\{\hat{\theta}_n \leq \theta\} = \frac{1}{2} + o(1), \quad P_{\theta,n}\{\hat{\theta}_n \geq \theta\} = \frac{1}{2} + o(1)$$

as $n \rightarrow \infty$, then $\hat{\theta}_n$ is called asymptotically median unbiased (AMU for short) for θ . Let $\mathcal{A}$ be a class of all the AMU estimators of θ .

Definition. If there exists an AMU estimator $\hat{\theta}_n^* = \hat{\theta}_n^*(\mathbf{X})$ such that for any $\hat{\theta}_n \in \mathcal{A}$, any $\theta \in \Theta$ and any $a > 0$

$$\begin{aligned} P_{\theta,n}\{|\hat{\theta}_n - \theta| > a\} &\geq P_{\theta,n}\{|\hat{\theta}_n^* - \theta| > a\}\{1 + o(1)\} \\ &=: B(a, \theta)\{1 + o(1)\} \end{aligned} \quad (1)$$

as $n \rightarrow \infty$, then $\hat{\theta}_n^*$ is called first order large-deviation efficient (LDE). If there exists an AMU estimator $\hat{\theta}_n^{**} = \hat{\theta}_n^{**}(\mathbf{X})$ such that for any $\hat{\theta}_n \in \mathcal{A}$, any $\theta \in \Theta$ and any $a > 0$

$$\frac{P_{\theta,n}\{|\hat{\theta}_n - \theta| > a\}}{B(a, \theta)} \geq 1 + \frac{b_1(a, \theta)}{n} + o\left(\frac{1}{n}\right) \quad (2)$$

as $n \rightarrow \infty$, and $\hat{\theta}_n^{**}$ attains the lower bound in (2), then $\hat{\theta}_n^{**}$ is second order LDE, where $b_1(a, \theta)$ is certain constant.

In order to discuss the higher order LDE, it is necessary to get the lower bound for the two-sided tail probability $P_{\theta,n}\{|\hat{\theta}_n - \theta| > a\}$ of $\hat{\theta}_n \in \mathcal{A}$.

3 First and second order LDEs

Let θ_0 be any fixed in Θ . Then we consider a problem of testing the hypothesis $H : \theta = \theta_0 + a$, against the alternative $K : \theta = \theta_0$, where $a > 0$. Let $\phi^*(\mathbf{X})$ be the most powerful (MP) test of level $1/2 + o(1)$ as $n \rightarrow \infty$. Putting

$$A_{\hat{\theta}_n} := \{\mathbf{x} | \hat{\theta}_n(\mathbf{x}) \leq \theta_0 + a\},$$

we see that the indicator $\chi_{A_{\hat{\theta}_n}}(\mathbf{x})$ is a test of level $1/2 + o(1)$ as $n \rightarrow \infty$, where $\mathbf{x} := (x_1, \dots, x_n)$. Since

$$E_{\theta_0}(\phi^*) \geq E_{\theta_0}[\chi_{A_{\hat{\theta}_n}}] = P_{\theta_0,n}\{\hat{\theta}_n \leq \theta_0 + a\}$$

for large n , it follows that

$$P_{\theta_0, n}\{\hat{\theta}_n - \theta_0 > a\} \geq 1 - E_{\theta_0}(\phi^*). \quad (3)$$

In order to obtain the lower bound, *i.e.*, the right-hand side of (3), it is seen from the fundamental lemma of Neyman-Pearson that a test with the rejection region of type

$$\bar{Z}(\theta_0) := \frac{1}{n} \sum_{j=1}^n Z_j(\theta_0) > c \quad (4)$$

is MP, where

$$Z_j(\theta_0) := \log(f(X_j, \theta_0)/f(X_j, \theta_0 + a)) \quad (j = 1, \dots, n)$$

and c is a constant chosen such that the asymptotic level of the test is $1/2 + o(1)$. Letting $\mu := E_{\theta_0+a}[Z_1(\theta_0)]$ and $\sigma^2 := V_{\theta_0+a}(Z_1(\theta_0))$, we put

$$W_n := \frac{\sqrt{n}}{\sigma} \{\bar{Z}(\theta_0) - \mu\}.$$

Since the MP test with the rejection region (3.2) is of asymptotic level $1/2$, *i.e.*

$$P_{\theta_0+a, n} \left\{ W_n \leq \frac{\sqrt{n}}{\sigma} (c - \mu) \right\} = \frac{1}{2} + o(1)$$

as $n \rightarrow \infty$, it follows from the Cornish-Fisher expansion that

$$\begin{aligned} \frac{\sqrt{n}}{\sigma} (c - \mu) &= u_{1/2} + \frac{1}{6} \kappa_{3, \theta_0+a}(W_n) (u_{1/2}^2 - 1) + O\left(\frac{1}{n}\right) \\ &= u_{1/2} + \frac{1}{6\sigma^3 \sqrt{n}} \kappa_{3, \theta_0+a}(Z_1(\theta_0)) (u_{1/2}^2 - 1) + O\left(\frac{1}{n}\right) = u_{1/2} + o(1), \end{aligned} \quad (5)$$

where $u_{1/2}$ is the upper 50 percentile of the standard normal distribution $N(0, 1)$, and $\kappa_{3, \theta_0+a}(W_n)$ and $\kappa_{3, \theta_0+a}(Z_1(\theta_0))$ are the third cumulants of W_n and $Z_1(\theta_0)$, respectively, under the hypothesis $H : \theta = \theta_0 + a$. Since $u_{1/2} = 0$, we have from (3.3)

$$c = \mu + o(1/\sqrt{n})$$

as $n \rightarrow \infty$. Note that $Z_1, \dots, Z_n$ are i.i.d. Since

$$E_{\theta_0}(\phi^*) = P_{\theta_0}\{\bar{Z}(\theta_0) > c\},$$

it follows from (3) that for large n

$$P_{\theta_0, n}\{\hat{\theta}_n - \theta_0 > a\} \geq 1 - P_{\theta_0}\{\bar{Z}(\theta_0) > c\} \quad (6)$$

for $a > 0$. In order to obtain the asymptotic expansion of the tail probability of $\bar{Z}(\theta_0)$ in (3.4), we use the saddlepoint approximation (Jensen (1995), Barndorff-Nielsen and Cox (1989)). Let

$$M(t) := E_{\theta_0}[\exp\{tZ_1(\theta_0)\}], \quad K(t) := \log M(t)$$

for all t in some open interval involving the origin, that is, they are the moment generating function (m.g.f.) and the cumulant generating function (c.g.f.) of $Z_1(\theta_0)$, respectively. Let $\hat{t}(a)$ be a solution of t of the equation $K'(t) = \mu$. In a similar way to the case $a > 0$, we have a lower bound in the case $a < 0$. Indeed, in the case when $a < 0$, we consider the indicator $\chi_{\hat{A}_{\theta_0}^c}(\mathbf{x})$ as a test of level $1/2 + o(1)$ in the same hypothesis testing problem as the case $a > 0$. Then we have for large n

$$P_{\theta_0, n}\{\hat{\theta}_n - \theta_0 \leq a\} \geq 1 - E_{\theta_0}(\phi^*),$$

where ϕ^* is the MP test of level $1/2 + o(1)$, which implies

$$P_{\theta_0, n} \{ \hat{\theta}_n - \theta_0 \leq a \} \geq 1 - P_{\theta_0, n} \{ \bar{Z}(\theta_0) > c \} \quad (7)$$

for $a < 0$.

From (3.4) and the saddlepoint approximation we have the following.

Theorem 1 *For any AMU estimator $\hat{\theta}_n$ of θ and any $a > 0$, it holds that for large n*

$$P_{\theta, n} \{ \hat{\theta}_n - \theta > a \} \geq \frac{1}{\lambda} M^n(\hat{t}) e^{-n\mu\hat{t}} \left[B_0(\lambda) + \frac{\text{sgn}(\hat{t}) \zeta_3(\hat{t})}{\sqrt{n}} \frac{1}{6} B_3(\lambda) \right. \\ \left. + \frac{1}{n} \left\{ \frac{\zeta_4(\hat{t})}{24} B_4(\lambda) + \frac{\zeta_3^2(\hat{t})}{72} B_6(\lambda) \right\} + O\left(\frac{1}{n^2}\right) \right],$$

where $\lambda = \sqrt{n} |\hat{t}| \sqrt{K''(\hat{t})}$, $\zeta_3(t) := \kappa_{3, \theta+a}(Z_1) / \{K''(t)\}^{3/2}$, and $\zeta_4(t) := \kappa_{4, \theta+a}(Z_1) / \{K''(t)\}^2$ with third and fourth cummulants $\kappa_{3, \theta+a}(Z_1)$ and $\kappa_{4, \theta+a}(Z_1)$ of $Z_1 = Z_1(\theta)$, and

$$B_0(\lambda) := \lambda e^{\lambda^2/2} \{1 - \Phi(\lambda)\} \approx \frac{1}{\sqrt{2\pi}} \quad (\lambda \rightarrow \infty), \\ B_3(\lambda) := - \left\{ \lambda^3 B_0(\lambda) - \frac{1}{\sqrt{2\pi}} (\lambda^3 - \lambda) \right\} \approx - \frac{3}{\sqrt{2\pi}\lambda} \quad (\lambda \rightarrow \infty), \\ B_4(\lambda) := \lambda^4 B_0(\lambda) - \frac{1}{\sqrt{2\pi}} (\lambda^4 - \lambda^2) \approx \frac{3}{\sqrt{2\pi}} \quad (\lambda \rightarrow \infty), \\ B_6(\lambda) := \lambda^6 B_0(\lambda) - \frac{1}{\sqrt{2\pi}} (\lambda^6 - \lambda^4 + 3\lambda^2) \approx - \frac{15}{\sqrt{2\pi}} \quad (\lambda \rightarrow \infty).$$

The proof is straightforward from (2.2.6) and section 2.1 in Jensen (1995). For the case $a < 0$, from (3.5) we also obtain a similar lower bound to (3.6) for the tail probability $P_{\theta, n} \{ \hat{\theta}_n - \theta < a \}$. If there exists an AMU estimator attaining the lower bound for the two-sided tail probability $P_{\theta, n} \{ |\hat{\theta}_n - \theta| > a \}$ up to the order $o(1/n)$, then it is second order LDE.

4 Examples

We shall apply the theorem to the case of a location parameter of the normal and two-sided exponential distributions. Without loss of generality, we assume that $\theta_0 = 0$ in the case.

Example 1. Suppose that $X_1, X_2, \dots, X_n, \dots$ is a sequence of i.i.d. random variables with the normal distribution $N(\theta, 1)$. Then, for $Z_1 = Z_1(0)$

$$\mu := E_a(Z_1) = E_a \left[\frac{a^2}{2} - aX_1 \right] = -\frac{a^2}{2}.$$

Since the m.g.f. of $Z_1(0)$ is $M(t) = \exp \{ (a^2/2)(t + t^2) \}$, it follows that the first derivative of the c.g.f. $K_1(t)$ is $K'(t) = (a^2/2)(1 + 2t)$, hence $\hat{t} = -1$. Since $\kappa_{3, a}(Z_1) = \kappa_{4, a}(Z_1) = 0$, it follows that $\zeta_3(t) = \zeta_4(t) = 0$. From (3.5) and (3.6) we have for

any fixed $a > 0$

$$\frac{P_{\theta,n}\{|\hat{\theta}_n - \theta| > a\}}{2\{1 - \Phi(\sqrt{na})\}} \geq 1 + O\left(\frac{1}{n^2}\right)$$

as $n \rightarrow \infty$. Letting $\bar{X} := (1/n) \sum_{i=1}^n X_i$, we have $P_{\theta,n}\{|\bar{X} - \theta| > a\} = 2\{1 - \Phi(\sqrt{na})\}$. Since $\bar{X}$ attains the lower bound, it is second order LDE.

Example 2. Suppose that $X_1, X_2, \dots, X_n, \dots$ is a sequence of i.i.d. random variables according to the two-sided exponential distribution with a p.d.f.

$$f(x, \theta) = \frac{1}{2} e^{-|x - \theta|} \quad (x \in \mathbb{R}^1; \theta \in \mathbb{R}^1).$$

Then the m.g.f. of $Z_1 = Z_1(0)$ is

$$M(t) := \frac{(t+1)e^{at} + te^{-a(t+1)}}{2t+1},$$

and

$$\mu = E_a(Z_1) = E_a[|X_1 - a| - |X_1|] = 1 - a - e^{-a},$$

and

$$K'(t) = \frac{\{a(t+1) + 1\}e^{at} - (at-1)e^{-a(t+1)}}{(t+1)e^{at} + te^{-a(t+1)}} - \frac{2}{2t+1},$$

hence $\hat{t} = -1$. Since

$$K''(-1) = 3 - 2(2a+1)e^{-a} - e^{-2a} = a^2 + O(a^3),$$

$$\kappa_{3,a}(Z_1) = -2\{e^{-3a} + 3(2a+1)e^{-2a} + 3(2a^2+2a+1)e^{-a} - 7\} = \frac{7}{4}a^4 + O(a^5),$$

$$\kappa_{4,a}(Z_1) = a^4 + O(a^5),$$

it follows that for a small $a > 0$

$$\zeta_3(-1) = \frac{\kappa_{3,a}(Z_1)}{\{K''(-1)\}^{3/2}} = \frac{7}{4}a + O(a^2), \quad (8)$$

$$\zeta_4(-1) = \frac{\kappa_{4,a}(Z_1)}{\{K''(-1)\}^2} = 1 + O(a). \quad (9)$$

Letting

$$\lambda = \sqrt{nK''(-1)} = a\sqrt{n} + O(\sqrt{a^3n})$$

for large n and

$$\mu = 1 - a - e^{-a} = -\frac{a^2}{2} + O(a^3)$$

in the Theorem, we have for large n

$$\begin{aligned} & P_{\theta,n}\{\hat{\theta}_n - \theta > a\} \\ & \geq \frac{1}{\lambda} e^{n(1-a-e^{-a})} \left[\lambda e^{\lambda^2/2} \{1 - \Phi(\lambda)\} + \frac{\zeta_3(-1)}{2\sqrt{2\pi}\sqrt{n}\lambda} + \frac{3\zeta_4(-1) - 5\zeta_3^2(-1)}{24\sqrt{2\pi}n} + O\left(\frac{1}{n^2}\right) \right], \end{aligned} \quad (10)$$

where $\zeta_3(-1)$ and $\zeta_4(-1)$ are given by (8) and (9), respectively. Using Mills' ratio, i.e.

$$1 - \Phi(x) = \phi(x) \left\{ \frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{3 \cdot 5}{x^7} + \dots + \frac{(-1)^k}{x^{2k+1}} \cdot \frac{(2k-1)!}{2^{k-1}(k-1)!} + O\left(\frac{1}{x^{2k+3}}\right) \right\}$$

as $x \rightarrow \infty$, we obtain for a fixed small $a > 0$

$$\begin{aligned} e^{n(1-a-e^{-a})+(\lambda^2/2)}\{1-\Phi(\lambda)\} &= \frac{1}{\sqrt{2\pi}}e^{n(1-a-e^{-a})}\left\{\frac{1}{\lambda}-O\left(\frac{1}{\lambda^3}\right)\right\} \\ &= \frac{1}{\sqrt{2\pi}}e^{n(-\frac{a^2}{2}+O(a^3))}\left\{\frac{1}{\sqrt{n}(a+O(a^2))}-O\left(\frac{1}{(\sqrt{n}(a+O(a^2)))^3}\right)\right\} \\ &= 1-\Phi(\sqrt{n}(a+O(a^2))) \end{aligned} \quad (11)$$

as $n \rightarrow \infty$. Then it follows from (4.1) to (4.5) and (3.5) that for a fixed small $a > 0$

$$\begin{aligned} \frac{P_{\theta,n}\{|\hat{\theta}_n - \theta| > a\}}{2\{1-\Phi(\sqrt{n}(a+O(a^2)))\}} &\geq 1 + \frac{\zeta_3(-1)}{2n(a+O(a^2))} + \frac{3\zeta_4(-1) - 5\zeta_3^2(-1)}{8n} + O\left(\frac{1}{n^2}\right) \\ &\geq 1 + \frac{5}{4n} + O\left(\frac{a}{n}\right) \end{aligned} \quad (12)$$

as $n \rightarrow \infty$, which gives the second order large-deviation lower bound.

Letting $n = 2k - 1$ and $S_a := \#\{i | X_i > a\}$ for $a > 0$, we see that S_a is distributed as a binomial distribution $B(n, p)$, where $p = (1/2)e^{-a}$. Then the maximum likelihood estimator of θ is the sample median $X_{(k)}$, and, by the saddlepoint approximation in the binomial case in Jensen (1995), page 43, the tail probability of $X_{(k)}$ is expanded as

$$\begin{aligned} P_{\theta,n}\{X_{(k)} - \theta > a\} &= P\left\{\frac{1}{n}S_a \geq \frac{k}{n}\right\} \\ &= \frac{p^k q^{n-k}}{1 - \left(\frac{\hat{q}p}{\hat{p}q}\right)^{\text{sgn}(\hat{p}-p)}} \frac{1}{\hat{p}^{k+(1/2)} \hat{q}^{n-k+(1/2)}} \left\{B_0(\lambda) + \frac{a_{\hat{p}}}{\sqrt{n}}B_1(\lambda) \right. \\ &\quad \left. + \frac{(1-2\hat{p})\text{sgn}(\hat{p}-p)}{6\sqrt{n}\hat{p}\hat{q}}B_3(\lambda) + O\left(\frac{1}{n^2}\right)\right\} \quad (n \rightarrow \infty) \end{aligned} \quad (13)$$

for a fixed $a > 0$, where $q := 1 - p$, $\hat{p} := k/n$, $\hat{q} := 1 - \hat{p}$, and

$$\begin{aligned} \hat{t} &:= \log \frac{\hat{p}q}{\hat{q}p}, \quad \hat{\sigma}^2(\hat{t}) = \hat{p}\hat{q}, \\ \gamma_{\hat{t}} &= \left(\frac{\hat{q}p}{\hat{p}q}\right)^{\text{sgn}(\hat{p}-p)} \left/ \left\{1 - \left(\frac{\hat{q}p}{\hat{p}q}\right)^{\text{sgn}(\hat{p}-p)}\right\}\right. = \begin{cases} \frac{\hat{p}\hat{q}}{\hat{p}-\hat{p}} & \text{for } \hat{p} > p, \\ \frac{\hat{p}\hat{q}}{\hat{p}-\hat{p}} & \text{for } \hat{p} < p, \end{cases} \\ a_{\hat{p}} &:= \frac{1}{\sigma(\hat{t})} \left(\frac{1}{|\hat{t}|} - \gamma_{\hat{t}}\right), \quad \lambda = \sqrt{n} \left| \log \frac{\hat{p}q}{\hat{q}p} \right| \sqrt{\hat{p}\hat{q}}, \\ B_1(\lambda) &= -\lambda \left\{B_0(\lambda) - (2\pi)^{-1/2}\right\} \approx \left(\sqrt{2\pi}\lambda\right)^{-1} \quad (\lambda \rightarrow \infty). \end{aligned}$$

Here we have

$$\begin{aligned}
 \log \left(\frac{p}{\hat{p}} \right)^k \left(\frac{q}{\hat{q}} \right)^{n-k} &= k \log \frac{p}{\hat{p}} + (n-k) \log \frac{q}{\hat{q}} \\
 &= n \left\{ \frac{k}{n} \log \frac{p}{\hat{p}} + \left(1 - \frac{k}{n} \right) \log \frac{q}{\hat{q}} \right\} \\
 &= -n \left\{ \hat{p} \log \left(1 + \frac{\hat{p}-p}{p} \right) + \hat{q} \log \left(1 - \frac{\hat{p}-p}{q} \right) \right\} \\
 &= -n \left\{ \frac{(\hat{p}-p)^2}{pq} - \frac{(\hat{p}-p)^2}{2pq} + O_p((\hat{p}-p)^3) \right\} \\
 &= -\frac{n(\hat{p}-p)^2}{2pq} + O_p\left(\frac{1}{\sqrt{n}}\right). \tag{14}
 \end{aligned}$$

Since

$$1 - \frac{\hat{q}p}{\hat{p}q} = \frac{\hat{p}-p}{\hat{p}q},$$

it follows from (13) and (14) that for $\hat{p} > p$

$$\begin{aligned}
 &P_{\theta,n}\{X_{(k)} - \theta > a\} \\
 &= \frac{1}{(\hat{p}-p)/(\hat{p}q)} \frac{1}{\sqrt{n\hat{p}\hat{q}}} \left[\exp \left\{ -\frac{n(\hat{p}-p)^2}{2pq} + O_p\left(\frac{1}{\sqrt{n}}\right) \right\} \right] \\
 &\quad \cdot \left\{ B_0(\lambda) + \frac{a_{\hat{p}}}{\sqrt{n}} B_1(\lambda) + \frac{(1-2\hat{p})\text{sgn}(\hat{p}-p)}{6\sqrt{n\hat{p}\hat{q}}} B_3(\lambda) + O_p\left(\frac{1}{n^2}\right) \right\} \quad (n \rightarrow \infty). \tag{15}
 \end{aligned}$$

Since for large n

$$\begin{aligned}
 \log \frac{\hat{p}q}{\hat{q}p} &= \log \left(1 + \frac{\hat{p}-p}{p} \right) - \log \left(1 - \frac{\hat{p}-p}{q} \right) \\
 &= \frac{\hat{p}-p}{pq} + O_p\left(\frac{1}{n^2}\right), \\
 \frac{\hat{p}-p}{\hat{p}q} &= \frac{\hat{p}-p}{pq} + O_p\left(\frac{1}{n^2}\right),
 \end{aligned}$$

it follows that

$$\lambda = \sqrt{n\hat{p}\hat{q}} \left\{ \frac{|\hat{p}-p|}{pq} + O_p\left(\frac{1}{n^2}\right) \right\}. \tag{16}$$

Since for large n

$$\begin{aligned}
 \lambda^2 &= \frac{n(\hat{p}-p)^2}{pq} \frac{\hat{p}\hat{q}}{pq} + O_p\left(\frac{1}{n\sqrt{n}}\right) \\
 &= \frac{n(\hat{p}-p)^2}{pq} \left(1 + \frac{\hat{p}-p}{p} \right) \left(1 - \frac{\hat{p}-p}{q} \right) + O_p\left(\frac{1}{n\sqrt{n}}\right) \\
 &= \frac{n(\hat{p}-p)^2}{pq} \left\{ 1 + O_p\left(\frac{1}{\sqrt{n}}\right) \right\},
 \end{aligned}$$

we obtain from (15) and (16) that

$$P_{\theta,n}\{X_{(k)} - \theta > a\} = \frac{1}{\lambda} e^{-\lambda^2/2} \left\{ B_0(\lambda) + \frac{a_{\hat{p}}}{\sqrt{n}} B_1(\lambda) + \frac{1-2\hat{p}}{6\sqrt{n\hat{p}\hat{q}}} \left(B_3(\lambda) + \frac{a_{\hat{p}}}{\sqrt{n}} B_4(\lambda) \right) + O\left(\frac{1}{n\sqrt{n}}\right) \right\}.$$

Since $p = (1/2)e^{-a}$, we have for a fixed small $a > 0$

$$\begin{aligned} \frac{1}{\sqrt{\hat{p}\hat{q}}} &= \frac{1}{\sqrt{pq}} \sqrt{\frac{pq}{\hat{p}\hat{q}}} = \frac{1}{\sqrt{pq}} \left\{ \left(1 + \frac{\hat{p}-p}{p}\right) \left(1 + \frac{\hat{q}-q}{q}\right) \right\}^{1/2} \\ &= \frac{1}{\sqrt{pq}} \left\{ 1 + O_p\left(\frac{a}{\sqrt{n}}\right) \right\} \\ &= 2(1 + O(a^2)) \left(1 + O_p\left(\frac{a}{\sqrt{n}}\right)\right) \end{aligned} \quad (17)$$

as $n \rightarrow \infty$, hence

$$\begin{aligned} a_{\hat{p}} &= \frac{1}{\sqrt{\hat{p}\hat{q}}} \left\{ \frac{pq}{\hat{p}-p} \left(1 - \frac{(p-q)(\hat{p}-p)}{2pq}\right) - \frac{p\hat{q}}{\hat{p}-p} + O_p\left(\frac{1}{\sqrt{n}}\right) \right\} \\ &= \frac{1}{2\sqrt{\hat{p}\hat{q}}} + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &= 1 + O(a^2) + O_p\left(\frac{a}{\sqrt{n}}\right). \end{aligned} \quad (18)$$

Since

$$\hat{p} = p + O_p\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{2}(1 - a) + O(a^2) + O_p\left(\frac{1}{\sqrt{n}}\right),$$

it follows from (15), (17) and (18) that for $\hat{p} > p$ and a fixed small $a > 0$

$$\frac{P_{\theta,n}\{X_{(k)} - \theta > a\}}{1 - \Phi(\lambda)} = 1 + \frac{1}{n} \left(\frac{1}{a} - \frac{1}{2}\right) + o\left(\frac{1}{na}\right) \quad (n \rightarrow \infty). \quad (19)$$

Since $\hat{p} = k/n = (1 + (1/n))/2$ and $p = (1/2)e^{-a}$, we have

$$\begin{aligned} \hat{p} - p &= \frac{1}{2}(1 - e^{-a}) + \frac{1}{2n} = \frac{1}{2}(a + O(a^2)) + \frac{1}{2n} > 0, \\ \sqrt{pq} &= \frac{1}{2} \sqrt{e^{-a}(2 - e^{-a})} = \frac{1}{2} \{(1 - a + O(a^2))(1 + a + O(a^2))\}^{1/2} \\ &= \frac{1}{2} + O(a^2), \end{aligned}$$

hence from (16) and (17)

$$\begin{aligned} \lambda &= \sqrt{n\hat{p}\hat{q}} \left\{ \frac{\hat{p}-p}{pq} + O\left(\frac{1}{n^2}\right) \right\} = \sqrt{n} \left(a + O(a^2) + O\left(\frac{1}{n}\right) \right) \\ &= \sqrt{n} (a + O(a^2)). \end{aligned} \quad (20)$$

Since the p.d.f. of the two-sided exponential distribution is symmetric, from (19) and (20) we obtain for a fixed small $a > 0$

$$\frac{P_{\theta,n}\{|X_{(k)} - \theta| > a\}}{2\{1 - \Phi(\sqrt{n}(a + O(a^2)))\}} = 1 + \frac{1}{n} \left(\frac{1}{a} - \frac{1}{2}\right) + o\left(\frac{1}{na}\right), \quad (21)$$

as $n \rightarrow \infty$. Hence, it follows from (12) and (21) that the MLE $X_{(k)}$ is first order LDE, but not second order LDE, for a small a .

5 Remarks

In the previous sections, we discuss the large-deviation efficiency up to the second order using the saddlepoint approximation for the tail probability of estimators. The approach is different from that of Fu (1982) where the second order efficiency of estimators is considered by the asymptotic expansion of the amount of K-L information in the lower bound for the tail probability of consistent estimator given by Bahadur (1971). A similar discussion for an exponential family of distributions to the above is also done by Akahira (2004).

References

- 1] Akahira, M. (1995). Statistical inference and large-deviation probability. (In Japanese). *Mathematical Science No. 380*, 56–62.
- 2] Akahira, M. (2004). Second order large-deviation efficiency of estimators for an exponential family of distributions. In preparation.
- 3] Akahira, M. and Takeuchi, K. (1981). *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*. Lecture Notes in Statistics **7**, Springer, New York.
- 4] Bahadur (1971). *Some Limit Theorems in Statistics*. Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.
- 5] Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman & Hall, London.
- 6] Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.
- 7] Fu, J. C. (1982). Large sample point estimation: A large deviation theory approach. *Ann. Statist.*, **10**, 762–771.
- 8] Ghosh, J. K. (1994). *Higher Order Asymptotics*. NSF-CBMS Regional Conference Series Probab. and Statist., **4**, Inst. of Math. Statist., Hayward, California.
- 9] Jensen, J. L. (1995). *Saddlepoint Approximations*, Clarendon Press, Oxford.
- 0] Pfanzagl, J. and Wefelmeyer, W. (1985). *Asymptotic expansions for General Statistical Models*. Lecture Notes in Statistics **31**, Springer, Berlin.

Data Augmentation and Gibbs Sampling for Regression Models of Small Counts

Sylvia Frühwirth-Schnatter and Helga Wagner

Johannes Kepler Universität Linz

Department of Applied Statistics and Econometrics

E-mail: sylvia.fruehwirth-schnatter@jku.at; helga.wagner@jku.at

Abstract: In this article we consider Bayesian analysis of Poisson regression models. Estimation is carried out within a Bayesian framework using data augmentation and MCMC methods. We suggest a new MCMC sampler, which possesses a Gibbs transition kernel, where we draw from full conditional distributions belonging to standard distribution families, only. This Gibbs sampler is applied to a standard Poisson regression model and to a Poisson regression models dealing with overdispersion.

Key words: Poisson count data, data augmentation, Gibbs sampling, overdispersion.

1 Introduction

Applied statisticians commonly have to deal with count data, recording for instance the number of road accidents or disease occurrences. Such data are necessarily non-negative integers and it is often appropriate to assume that the observed counts follow a Poisson distribution. The Poisson regression model, discussed for instance in [12] in the framework of generalized linear models, is an important tool for analyzing the effect of covariates on count data. The basic Poisson regression model has been modified in a number of ways. To account for the dependency likely to be present in sequences of counts data, the covariates may depend on past observations, see for instance [19]. A couple of extensions deal with overdispersion due to omitted covariates, we mention in particular mixtures of Poisson regression models [17, 9], Poisson regression models with additive random effects [1], panel Poisson regression models with random effects [4], and mixtures of Poisson regression models with random effects [11].

In this paper we consider Bayesian estimation of Poisson regression models, using data augmentation as in [15] and Markov chain Monte Carlo (MCMC) methods, as illustrated first by Zeger and Karim [18] for generalized linear models with random effects. Since this seminal paper, a number of authors have contributed to MCMC estimation of regression models for count data. We mention here in particular Albert [2] for Poisson random-effects models, Chib *et al.* [4] for panel count data models with multiple random effects, Lenk and DeSarbo [11] for mixtures of Poisson models with random effects, and Hurn *et al.* [9] for mixtures of Poisson regression models.

A major difficulties with any of the existing MCMC approaches is that practical implementation requires the use of a Metropolis-Hastings algorithm at least for part of the unknown parameter vector, which in turns make it necessary to define suitable proposal densities. The present article discusses a new method of data augmentation, and straightforward Gibbs sampling, put forward by Frühwirth-Schnatter and Wagner [7], in the context of Poisson regression models. Our main result is to show that a Poisson regression model may be regarded as a partially Gaussian regression model in the sense of [14], by conditioning on two sequences of suitably chosen artificially missing data. The first sequence are the unobserved inter-arrival times of a suitably chosen Poisson process. This eliminates the non-linearity of the Poisson regression model, and leads to a linear regression model with non-normal errors, that follow a log exponential distribution with mean 1. The log exponential distribution is then approximated by a mixture of normal distributions in a similar way as in [10] and [5]. We introduce the component indicator of this normal mixture as a second sequence of missing data. By conditioning on both sequences, a Gaussian regression model results. Based on this useful result, we will show that straightforward Gibbs sampling of all regression parameters, and all missing data is possible, requiring only random draws from standard distributions such as multivariate normals, inverse Gamma, exponential and discrete distributions with a few categories.

The rest of the paper is organized as follows. After a short review of the Poisson regression model in Section 2, we discuss in Section 3 in detail data augmentation for Poisson regression models, and implementation of our new Gibbs sampling scheme. Applications to standard regression models as well as to regression models dealing with overdispersion are considered in Section 4, whereas Section 5 concludes.

2 Poisson Regression Models

Let $\mathbf{y} = (y_1, \dots, y_N)$ be a collection of count data. In what follows, we assume that $y_i | \lambda_i$ follows a $\mathcal{P}(\lambda_i)$ distribution, where λ_i depends on an unknown model parameter $\boldsymbol{\vartheta}$ in the following way:

$$y_i | \lambda_i \sim \mathcal{P}(\lambda_i), \quad \lambda_i = \exp(\mathbf{z}_i \boldsymbol{\vartheta}), \quad (1)$$

where $\mathbf{z}_i$ is a row vector of regressors, including 1 for the intercept. In the present paper we will be interested in two special cases of this model. First, we consider

the standard Poisson regression model, where for each count variable y_i , we observe covariates $\mathbf{x}_i$, again including 1 for the intercept. Then in (1), $\mathbf{z}_i = \mathbf{x}_i$, and $\boldsymbol{\vartheta} = \boldsymbol{\beta}$ is a simple regression parameter. We will show below, that our data augmentation method leads to a normal regression model, where the whole regression parameter $\boldsymbol{\beta}$ could be sampled in one sweep from a normal distribution.

Second, we will consider a Poisson regression model dealing with overdispersion due to omitted covariates. A common way of dealing with this kind of overdispersion is the individual effects model introduced by [1], where the regression intercept varies between the units:

$$\lambda_i = \exp(\alpha_i + \mathbf{x}_i \boldsymbol{\beta}), \quad (2)$$

where $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$. Thus overdispersion is modelled on the same level as the linear predictor. Formally, this model may be written as a special case of (1), where $\boldsymbol{\vartheta} = (\alpha_1, \dots, \alpha_N, \boldsymbol{\beta})$, and

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{e}_i & \mathbf{x}_i \end{pmatrix},$$

where $\mathbf{e}_i$ is a $(1 \times N)$ row vector, containing only zeros, apart from column i , which is equal to 1. Marginally, this model is an infinite mixture of Poisson regression models with no closed form. Aitkin [1] suggested to approximate the marginal distribution by a mixture of Poisson regression models using Gaussian-Hermite quadrature. We will show, how data augmentation leads to a normal random-effects regression model, where the whole sequence $\boldsymbol{\vartheta} = (\alpha_1, \dots, \alpha_N, \boldsymbol{\beta})$ could be sampled simultaneously in an efficient manner.

3 A Bayesian Analysis

3.1 Prior and Posterior Distributions

We assume that the prior distribution $p(\boldsymbol{\vartheta}|\boldsymbol{\delta})$ of $\boldsymbol{\vartheta}$ follows a normal distribution, which is allowed to be indexed by an unknown hyper parameter $\boldsymbol{\delta}$. For the standard regression model $p(\boldsymbol{\beta}|\boldsymbol{\delta})$ typically is the normal prior $\mathcal{N}_d(\mathbf{b}_0, \mathbf{B}_0)$, with known hyperparameters $\mathbf{b}_0$ and $\mathbf{B}_0$. For a regression model with overdispersion, this prior is extended to:

$$p(\boldsymbol{\vartheta}|\boldsymbol{\delta}) = p(\boldsymbol{\beta}|\mathbf{b}_0, \mathbf{B}_0) \prod_{i=1}^N p(\alpha_i|\sigma_\alpha^2),$$

where $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$. The additional hyperparameter σ_α^2 may be known or unknown.

For both models, these assumptions are sufficient to derive the conditional posterior density $p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \mathbf{y})$ by Bayes' theorem, given all observations $\mathbf{y} = (y_1, \dots, y_N)$:

$$p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\boldsymbol{\delta}), \quad p(\mathbf{y}|\boldsymbol{\vartheta}) = \prod_{i=1}^N \frac{\exp(\mathbf{z}_i \boldsymbol{\vartheta})^{y_i}}{y_i!} \exp(-\exp(\mathbf{z}_i \boldsymbol{\vartheta})).$$

The resulting posterior density, however, in general does not belong to a density from a well-known distribution family. Markov chain Monte Carlo methods to sample from

the posterior distribution of a Poisson regression model were applied by Zeger and Karim [18], Albert [2], Chib *et al.* [4], Lenk and DeSarbo [11] and Hurn *et al.* [9], among many others. As mentioned in the introduction, any of these methods is based on Metropolis-Hastings sampling.

We are going to demonstrate in the following subsection, that the introduction of two sequences of artificially missing data within a data augmentation scheme leads to a conditional posterior distribution for $\boldsymbol{\vartheta}$ that, in contrast to $p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \mathbf{y})$, is a joint normal distribution, once we conditioned on the artificially missing data.

3.2 Data Augmentation

For each i , the distribution of $y_i|\lambda_i$ may be regarded as the distribution of the number of jumps of an unobserved Poisson process with intensity λ_i , having occurred in the time interval $[0,1]$. The first step of data augmentation introduces for each i , $i = 1, \dots, N$, the inter-arrival times τ_{ij} , $j = 1, \dots, (y_i + 1)$ of this Poisson process as missing data. From the basic properties of a Poisson process, the inter-arrival times τ_{ij} are known to follow the $\mathcal{E}(\lambda_i)$ -distribution:

$$\tau_{ij}|\boldsymbol{\vartheta} \sim \mathcal{E}(\lambda_i) = \frac{\mathcal{E}(1)}{\lambda_i}.$$

As $\lambda_i = \exp(\mathbf{z}_i\boldsymbol{\vartheta})$, this may be reformulated as following linear model:

$$\log \tau_{ij}|\boldsymbol{\vartheta} = -\mathbf{z}_i\boldsymbol{\vartheta} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \log(\mathcal{E}(1)). \quad (3)$$

Let $\boldsymbol{\tau} = \{\tau_{ij}, j = 1, \dots, (y_i + 1), i = 1, \dots, N\}$ denote the collection of all inter-arrival times. Our first data augmentation step introduces the inter-arrival times $\boldsymbol{\tau}$ as missing data, with two effects. First, the full-conditional posterior distribution $p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \boldsymbol{\tau}, \mathbf{y})$ of $\boldsymbol{\vartheta}$, where additionally to $\boldsymbol{\delta}$ and $\mathbf{y}$ the inter-arrival times $\boldsymbol{\tau}$ appear as conditioning argument, is independent of $\mathbf{y}$, $p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \boldsymbol{\tau}, \mathbf{y}) = p(\boldsymbol{\vartheta}|\boldsymbol{\delta}, \boldsymbol{\tau})$. Second, conditional on $\boldsymbol{\tau}$, we are dealing with model (3), which is non-normal, but where the mean of the observation equation is linear in the unknown model parameters $\boldsymbol{\vartheta}$. Thus, the first augmentation steps eliminates the non-linearity of the Poisson regression model, the non-normality of the error term, however, remains.

It is important to realize that the error term in (3) follows a $\log \mathcal{E}(1)$ -distribution which is independent of any unknown model parameter. To obtain a model that is conditionally Gaussian, we start by approximating the non-normal density of $\varepsilon_{ij} \sim \log(\mathcal{E}(1))$ by a normal mixture of 5 components with parameters m_r and s_r for the r -th component:

$$p(\varepsilon_{ij}) = \exp\{\varepsilon_{ij} - e^{\varepsilon_{ij}}\} \approx \sum_{r=1}^5 w_r f_{\mathcal{N}}(\varepsilon_{ij}; m_r, s_r^2). \quad (4)$$

This idea is influenced by the related articles of Kim *et al.* [10] and Chib *et al.* [5], who used a normal mixture approximation of the density of a $\log \chi^2$ -distribution in the context of stochastic volatility models. The appropriate parameters (w_r, m_r, s_r^2) , $r = 1, \dots, 5$, however, are different for our problem and are tabulated in Table 3.2 for 5 components, a number that we found to be sufficiently large in practice. The para-

r	1	2	3	4	5
w_r	0.2924	0.2599	0.2480	0.1525	0.0472
m_r	0.0982	-1.5320	-0.7433	0.8303	-3.1428
s_r^2	0.2401	1.1872	0.3782	0.1920	3.2375

Table 1: Normal mixture approximation of the density of the $\log \mathcal{E}(1)$ -distribution (5 components)

parameters of the mixture approximation were determined by minimizing the Kullback-Leibler distance.

Following [10] and [5], the mixture distribution (4) is regarded as the marginal distribution of a problem where additional to ε_{ij} the component indicators r_{ij} are observed. The second step of our data augmentation scheme introduces for each ε_{ij} the latent component indicator r_{ij} as missing data. Let $\mathbf{R} = \{r_{ij}, j = 1, \dots, (y_i + 1), i = 1, \dots, N\}$ denote the collection of all component indicators r_{ij} . Conditional on $\boldsymbol{\tau}$ and $\mathbf{R}$ the Poisson regression model (1) reduces to a Gaussian regression model with heteroscedastic errors with known variance:

$$\log \tau_{ij} | \boldsymbol{\vartheta}, r_{ij} = -\mathbf{z}_i \boldsymbol{\vartheta} + m_{r_{ij}} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}\left(0, s_{r_{ij}}^2\right). \quad (5)$$

For such a model the conditional posterior $p(\boldsymbol{\vartheta} | \boldsymbol{\delta}, \boldsymbol{\tau}, \mathbf{R}, \mathbf{y})$, is a multivariate normal density, which is easy to sample from. This result is the basis for our new two-block Gibbs sampler, that will be described in the next subsection.

3.3 The Basic Two-block Gibbs Sampler

When the hyperparameter $\boldsymbol{\delta}$ of the prior $p(\boldsymbol{\vartheta} | \boldsymbol{\delta})$ is assumed to be known, a two-block Gibbs sampler results, if data augmentation as described in the previous section is applied. Select a starting value for the component indicators $\mathbf{R} = \{r_{ij}, j = 1, \dots, y_i + 1, i = 1, \dots, N\}$, and the inter-arrival times $\boldsymbol{\tau} = \{\tau_{ij}, j = 1, \dots, y_i + 1, i = 1, \dots, N\}$ and repeat the following steps:

- a) Multi-move sampling of $\boldsymbol{\vartheta}$ conditional on knowing $\boldsymbol{\tau}$, $\mathbf{R}$, and $\mathbf{y}$, based the normal regression model (5).
- b) Sampling of the inter-arrival times $\boldsymbol{\tau}$ and $\mathbf{R}$ conditional on knowing $\boldsymbol{\vartheta}$ and $\mathbf{y}$. For $i = 1, \dots, N$ run through the following steps (b1) to (b3) with $\lambda_i = \exp(\mathbf{z}_i \boldsymbol{\vartheta})$ and $n = y_i$:
 - (b1) If $y_i > 0$, sample the order statistics $u_{t,(1)}, \dots, u_{t,(n)}$ of n uniformly distributed random variables, see e.g. [13] for details, and define the inter-arrival times τ_{ij} as their increments: $\tau_{ij} = u_{i,(j)} - u_{i,(j-1)}, j = 1, \dots, n$, where $u_{i,(0)} := 0$.
 - (b2) Sample the final arrival time as $\tau_{i,n+1} = 1 - \sum_{j=1}^n \tau_{ij} + \xi_i$, where $\xi_i \sim \mathcal{E}(\lambda_i)$.

- (b3) For each $j = 1, \dots, y_i + 1$, sample the component indicators r_{ij} conditional on τ_{ij} and ϑ from the following discrete density:

$$\Pr(r_{ij} = k | \tau_{ij}, \vartheta) \propto p(\tau_{ij} | r_{ij} = k, \vartheta) w_k,$$

where

$$\log p(\tau_{ij} | r_{ij} = k, \delta, \vartheta) \propto -\log s_k - \frac{1}{2} \left(\frac{\log \tau_{ij} + \mathbf{z}_i \vartheta - m_k}{s_k} \right)^2.$$

The quantities $(w_k, m_k, s_k^2), k = 1, \dots, 5$ are the parameters of the finite mixture approximation tabulated in Table 3.2.

Note that step (b) involves only draws from standard densities. Thus sampling scheme (a) and (b) is actually a Gibbs sampler without any tuning.

Comments

It is easy to verify the different sampling steps (b1) to (b3). The joint posterior $p(\mathbf{R}, \boldsymbol{\tau} | \mathbf{y}, \vartheta)$ is decomposed as:

$$p(\mathbf{R}, \boldsymbol{\tau} | \mathbf{y}, \vartheta) = p(\mathbf{R} | \boldsymbol{\tau}, \mathbf{y}, \vartheta) p(\boldsymbol{\tau} | \mathbf{y}, \vartheta).$$

The inter-arrival times $\{\tau_{ij}, j = 1, \dots, y_i + 1\}$ are independent for different time points i , given $\mathbf{y}$, and ϑ :

$$p(\boldsymbol{\tau} | \mathbf{y}, \vartheta) = \prod_{i=1}^N p(\tau_{i1}, \dots, \tau_{i, y_i}, \tau_{i, y_i+1} | y_i, \vartheta).$$

For fixed i , the inter-arrival times $\tau_{i1}, \dots, \tau_{i, n+1}$, where $n = y_i$, are stochastically dependent, and the joint distribution factorizes as:

$$\begin{aligned} p(\tau_{i1}, \dots, \tau_{in}, \tau_{i, n+1} | y_i = n, \vartheta) \\ = p(\tau_{i, n+1} | y_i = n, \vartheta, \tau_{i1}, \dots, \tau_{in}) p(\tau_{i1}, \dots, \tau_{in} | y_i = n). \end{aligned}$$

The first n inter-arrival times are independent of ϑ and the component indicator $\mathbf{R}$, and are determined only by the observed number of counts y_i . Due to well-known properties of a Poisson process, the n arrival times occurring in $[0, 1]$ are distributed as the order statistics of $\mathcal{U}[0, 1]$ -distributed random variables, and step (b1) follows immediately. The final inter-arrival time $\tau_{i, n+1}$ depends on the actual model parameters ϑ through the risk λ_i , but is also independent of the component indicator $\mathbf{R}$. Conditionally on $y_i = n$ and $\tau_{i1}, \dots, \tau_{in}$, the last arrival time $\tau_{i, n+1}$ has an exponential distribution with mean $1/\lambda_i$, truncated at $1 - \sum_{j=1}^n \tau_{ij}$, thus step (b2) follows.

The component indicators r_{ij} are mutually independent for different i as well as for different j , given $\boldsymbol{\tau}$, ϑ and $\mathbf{y}$:

$$p(\mathbf{R} | \boldsymbol{\tau}, \mathbf{y}, \vartheta) = \prod_{i=1}^N \prod_{j=1}^{y_i+1} p(r_{ij} | \tau_{ij}, \vartheta).$$

For i, j fixed, the posterior of each component indicator r_{ij} depends on the data only through τ_{ij} and on the model parameters ϑ only through the risk λ_i , thus step (b3) follows immediately.

Step (b1) could be used to sample starting values for $\tau_{i1}, \dots, \tau_{in}$ for each i , given the observed counts y_i . To obtain a starting value for $\tau_{i,n+1}$, we use (b2) and sample ξ_i from $\mathcal{E}(\lambda_i)$ with $\lambda_i = y_i$. For all i , where $y_i = 0$, λ_i can be set to a “small” value for λ_i , in our examples we used $\lambda_i = 0.1$. Starting values for each component indicator r_{ij} are obtained as random draws from 1 to 5.

4 Applications

4.1 Application to the Standard Poisson Regression Model

For the Poisson regression model $y_i \sim \mathcal{P}(\lambda_i)$, $\log \lambda_i = \mathbf{x}_i \boldsymbol{\beta}$, data augmentation by $\boldsymbol{\tau}$ and $\mathbf{R}$ as described above, leads to a normal regression model with $n_i = y_i + 1$ repeated measurements, and heteroscedastic errors with known variance. For each i , this model reads:

$$\log \tau_{ij} | \boldsymbol{\theta}, r_{ij} = -\mathbf{x}_i \boldsymbol{\beta} + m_{r_{ij}} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}\left(0, s_{r_{ij}}^2\right), \quad (6)$$

where $j = 1, \dots, y_i + 1$.

Under the normal prior $\boldsymbol{\beta} \sim \mathcal{N}_d(\mathbf{b}_0, \mathbf{B}_0)$, the conditional posterior $p(\boldsymbol{\beta} | \boldsymbol{\tau}, \mathbf{R}, \mathbf{y}) = p(\boldsymbol{\beta} | \boldsymbol{\tau}, \mathbf{R})$ is equal to the multivariate normal density $\mathcal{N}_d(\mathbf{b}_N, \mathbf{B}_N)$ with following posterior moments:

$$\mathbf{B}_N^{-1} = \mathbf{B}_0^{-1} + \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i w_i, \quad \mathbf{b}_N = \mathbf{B}_N (\mathbf{B}_0^{-1} \mathbf{b}_0 - \sum_{i=1}^N \mathbf{x}_i' \tilde{y}_i), \quad (7)$$

where

$$w_i = \sum_{j=1}^{y_i+1} \frac{1}{s_{r_{ij}}^2}, \quad \tilde{y}_i = \sum_{j=1}^{y_i+1} \frac{\log \tau_{ij} - m_{r_{ij}}}{s_{r_{ij}}^2}. \quad (8)$$

This result is derived in a straightforward manner from a Bayesian analysis of the normal regression model (6), see [20], among many other references.

4.1.1 Application to Road Safety Data

In this application we use the Gibbs sampler to analyze a data set provided by the Austrian Road Safety Board. These data are monthly counts of killed or injured pedestrians, aged 6-10 in Linz, which is the third largest town in Austria. The time period covered was 1987 to 2002. A legal intervention intended to increase road safety took place during the observation period. On October 1, 1994 an amendment increasing priority for pedestrians became effective: since then pedestrians who want to use a crosswalk have to be allowed crossing without risk. We model these data using a regression model with fixed seasonal effects s_i for the different months and the intervention effect is modelled as a level shift at the time point $t = t_{int}$ when legal amendments became effective. Our model parameters are an intercept β , the

Parameter	Gibbs Sampling			Maximum Likelihood	
	Mean	Std.dev	95%H.P.D. regions	Mean	Std.dev
β	0.7028	0.0756	[0.5501, 0.8466]	0.7177	0.0730
δ	-0.3553	0.1080	[-0.5629, -0.1379]	-0.3543	0.1077

Table 2: Parameter estimates for the regression model

seasonal dummies $s_1, \dots, s_{11}$ and the intervention effect δ , the covariate vector is of dimension 1×13 .

The Gibbs sampler described in Subsection 3.3 was run 12000 times with a burn in of 2000 runs. Table 2 reports point estimates as well as 95%-H.P.D. regions for the intercept and the intervention effect and compares them to the maximum likelihood estimates. The intervention effect is significantly negative. ML estimates and MCMC estimates are very similar which is also true for the seasonal effects.

Figure 1 shows the observed counts with the exponentiated estimated level including the intervention effect and pointwise 95% credible intervals and the seasonal pattern $\exp(s_i)$. The main feature of the seasonal pattern in the children series is a significant decrease in holiday months July and August.

Table 3 shows the parameter estimates obtained from a Gibbs sampler based on a five component mixture approximation in comparison to the maximum likelihood estimates. In the same table we study the effect of using less than five components in the mixture approximation to the distribution of ε_{ij} . Not surprisingly, we find a large improvement of choosing a mixture approximation with at least two components, rather than the one component normal approximation. The small differences between the different higher component estimates are only due to the Monte Carlo integration error.

4.2 Poisson Regression Models Dealing With Overdispersion

A common problem in Poisson regression models is the presence of overdispersion due to omitted covariates which may cause bias and loss of efficiency in estimating the remaining regression parameters, see for instance [6]. A simple, but useful way of dealing with this kind of overdispersion is to introduce individual effects $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$ into the linear predictor:

$$y_i \sim \mathcal{P}(\lambda_i), \quad \log \lambda_i = \alpha_i + \mathbf{x}_i \beta.$$

The Bayesian estimation methods discussed in Subsection 4.1 are easily extended to this more general setting. Data augmentations by τ and $\mathbf{R}$ as described above leads to a normal individual effects regression model with $n_i = y_i + 1$ repeated measurements, and heteroscedastic errors with known variance. For each i , this model reads:

$$\log \tau_{ij} | \alpha_i, \beta, r_{ij} = -\alpha_i - \mathbf{x}_i \beta + m_{r_{ij}} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, s_{r_{ij}}^2), \quad (9)$$

where $j = 1, \dots, y_i + 1$. The joint conditional posterior $p(\vartheta | \tau, \mathbf{R}, \mathbf{y})$ of $\vartheta = (\alpha_1, \dots, \alpha_N, \beta)$ could easily be derived as in Subsection 4.1, however, joint sampling from this ex-

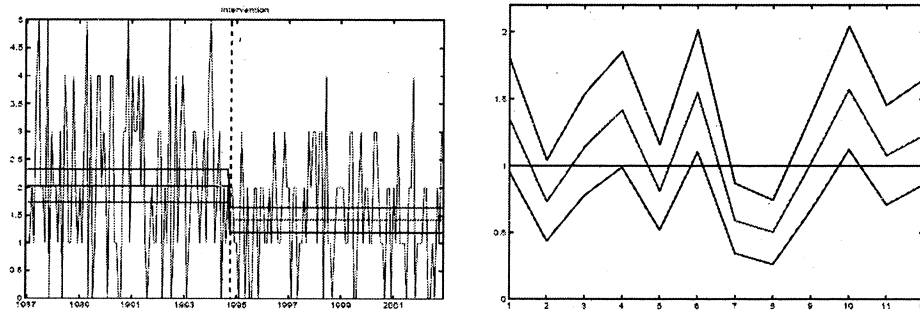


Figure 1. Counts of killed or injured children with estimated rate (posterior means, left) and seasonal pattern (posterior means, right) within 95% credible region.

Method		β		δ	
		Mean	Std.dev	Mean	Std.dev
Maximum Likelihood		0.7177	0.0730	-0.3543	0.1077
Gibbs Sampling	1 component	0.6938	0.0888	-0.4014	0.1307
	2 components	0.7096	0.0719	-0.3611	0.1055
	3 components	0.6978	0.0735	-0.3535	0.1070
	4 components	0.6971	0.0747	-0.3544	0.1087
	5 components	0.7028	0.0756	-0.3553	0.1080

Table 3: Comparison of parameter estimates for different methods.

tremely high-dimensional normal density through the Cholesky decomposition of the corresponding variance-covariance matrix is not very efficient. We discuss an efficient multi move sampler.

4.2.1 Efficient multi move sampling

It is possible to write the joint conditional posterior of $\boldsymbol{\vartheta} = (\alpha_1, \dots, \alpha_N, \boldsymbol{\beta})$ as

$$p(\boldsymbol{\vartheta}|\boldsymbol{\tau}, \mathbf{R}, \mathbf{y}) = p(\boldsymbol{\vartheta}|\boldsymbol{\tau}, \mathbf{R}) = p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R}) \prod_{i=1}^N p(\alpha_i|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{R}).$$

$p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R})$ is derived from a regression model, where the individual effects appearing in (9) are integrated out.

For each $i, i = 1, \dots, N$, all inter arrival times τ_{ij} , generated for a specific count observation y_i , for $j = 1, \dots, n_i$, where $n_i = (y_i + 1)$, share the same individual effect α_i and are independent only conditional on knowing α_i . The marginal model for $\log \boldsymbol{\tau}_i = (\log \tau_{i1} \dots \log \tau_{i,n_i})'$ given only $\boldsymbol{\beta}$, with the individual effect being integrated out, is

$$\log \boldsymbol{\tau}_i = -\mathbf{1}\mathbf{x}_i\boldsymbol{\beta} + m_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(\mathbf{0}, V_i), \quad (10)$$

where $m_i = (m_{r_{i1}} \dots m_{r_{i,n_i}})'$ and

$$V_i = \mathbf{1}\mathbf{1}'\sigma_\alpha^2 + \text{Diag}(s_{r_{i1}}^2, \dots, s_{r_{i,n_i}}^2).$$

From the marginal model (10) we obtain that the posterior $p(\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{R})$ is equal to a multivariate normal density $\mathcal{N}_d(\mathbf{b}_N, \mathbf{B}_N)$ with the posterior moments $\mathbf{b}_N$ and $\mathbf{B}_N$ given by

$$\mathbf{B}_N^{-1} = \mathbf{B}_0^{-1} + \sum_{i=1}^N \mathbf{x}_i' \mathbf{1}' V_i^{-1} \mathbf{1} \mathbf{x}_i, \quad \mathbf{b}_N = \mathbf{B}_N(\mathbf{B}_0^{-1} \mathbf{b}_0 - \sum_{i=1}^N \mathbf{x}_i' \mathbf{1}' V_i^{-1} (\log \boldsymbol{\tau}_i - m_i)). \quad (11)$$

For each $i = 1, \dots, N$, the conditional posterior $p(\alpha_i|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{R})$ is derived from regression model (9), with $\boldsymbol{\beta}$ considered to be known. From the results of Subsection 4.1, under the normal prior $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$, the conditional posterior $p(\alpha_i|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{R})$ is equal to a univariate normal density $\mathcal{N}(a_i(\boldsymbol{\beta}), A_i)$ with following posterior moments:

$$A_i^{-1} = \frac{1}{\sigma_\alpha^2} + \sum_{j=1}^{y_i+1} \frac{1}{s_{r_{ij}}^2}, \quad a_i(\boldsymbol{\beta}) = A_i \left(\sum_{j=1}^{y_i+1} \frac{\log \tau_{ij} + \mathbf{x}_i \boldsymbol{\beta} - m_{r_{ij}}}{s_{r_{ij}}^2} \right). \quad (12)$$

Thus step (a) in the Gibbs sampler is implemented by sampling $\boldsymbol{\beta}$ from $\mathcal{N}_d(\mathbf{b}_N, \mathbf{B}_N)$, with the moments given by (11), and then sampling α_i from $\mathcal{N}(a_i(\boldsymbol{\beta}), A_i)$ for $i = 1, \dots, N$.

4.2.2 Dealing with unknown hyperparameters

If the hyperparameter σ_α^2 is unknown, then an additional step (c) has to be added, to sample σ_α^2 from the inverted Gamma posterior distribution $\mathcal{G}^{-1}(c_N/2, C_N/2)$ with $c_N = c_0 + N$, and $C_N = C_0 + \sum_{i=1}^N \alpha_i^2$, under the inverted Gamma prior distribution $\sigma_\alpha^2 \sim \mathcal{G}^{-1}(c_0/2, C_0/2)$.

Parameter	low heterogeneity $\sigma_\alpha^2 = 0.0001$		high heterogeneity $\sigma_\alpha^2 = 1$	
	Mean	Std.dev	Mean	Std.dev
β_0	0.5391	0.1196	0.4450	0.1895
β_1	-0.3035	0.1107	-0.3339	0.1659
σ_α^2	0.0070	0.0110	1.0020	0.2037

Table 4: Parameter estimates for the heterogeneity model

4.2.3 Application to Simulated Data

For illustration, we consider two simulated data sets of size $N = 200$, generated according to the heterogeneity model

$$y_i \sim \mathcal{P}(\lambda_i), \quad \log \lambda_i = \alpha_i + \beta_0 + \beta_1 x_i,$$

where x_i corresponds to a linear trend and $\beta = (\beta_0, \beta_1) = (0.5, -0.3)$. The variance of the random effects is $\sigma_\alpha^2 = 0.0001$ for the first data set, which implies a low degree of unobserved heterogeneity, whereas $\sigma_\alpha^2 = 1$ for the second data set, causing a high degree of unobserved heterogeneity.

A two-block Gibbs sampler based on the five component mixture approximation was used to estimate the unknown parameters. Conditional on knowing σ_α^2 , we sample the whole parameter $\vartheta = (\alpha_1, \dots, \alpha_N, \beta)$ jointly within one move using the results of Subsection 1. Conditional on knowing ϑ , the variance σ_α^2 is sampled as described in Subsection 2. The Gibbs sampler was run 12000 times with a burn in of 2000 runs.

Estimation results are presented in Table 3, autocorrelation functions of the estimated values of β and σ_α^2 for both the low and high heterogeneity case are shown in Figure 2. As for Gaussian random-effects models, see for instance [8] and [16], the convergence behavior of Gibbs sampling is worse in the low heterogeneity case (σ_α^2 close to 0), than for the high heterogeneity case ($\sigma_\alpha^2 = 1$).

5 Concluding Remarks

The Gibbs sampler studied in this paper provides an important step toward operational MCMC estimation for a broad class of regression models for Poisson counts, as our sampler requires only draws from standard densities, without tuning of proposal densities. Gibbs sampling is feasible for most of the Poisson regression models suggested in the literature so far. The Gibbs sampler discussed in this paper for a standard regression model and a Poisson regression model with additive random effects is easily modified to deal with various extensions of the Poisson regression model, in particular with mixtures of Poisson regressions models [17, 9], panel Poisson regression models with random effects [4], and mixtures of Poisson regression models with random effects [11]. Space limits do not allow us to give all the details, which

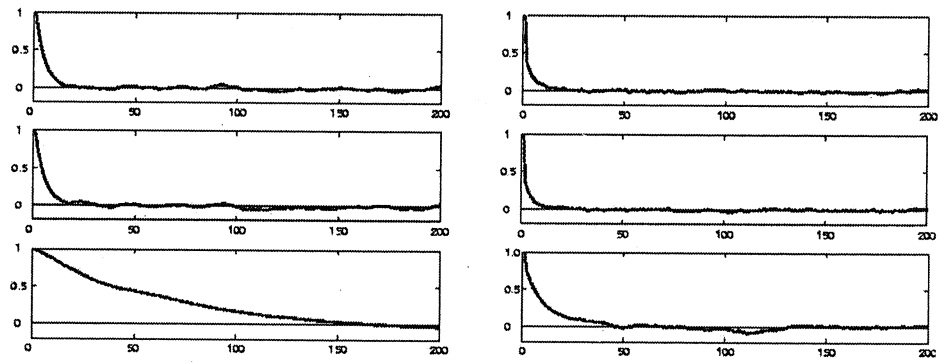


Figure 2. ACF of β_0 (above), β_1 (middle) and σ_α^2 (below); low heterogeneity case ($\sigma_\alpha^2=0.01$) left side, high heterogeneity case ($\sigma_\alpha^2=1$) right side.

are, however, pretty straightforward, and are worked out in detail for time-varying Poisson regression models in [7] .

Our new data augmentation scheme seems to be a promising step toward solving further issue in Bayesian estimation for generalized linear models based on the Poisson distribution. First, for random effects models some care must be exercised with respect to parameterization issues, as Gibbs sampling often leads to convergence problems, if σ_α^2 is close to 0. Such problems are well-known for Gaussian random-effects model, see for instance [8] and [16] . For Poisson count data parameterization issues are also addressed in [4] . Our new data augmentation may be helpful in this regard, as a non-centered parameterization similar to the one studied in [16] for Gaussian models, is easily available.

Finally, we would like to mention that our data augmentation scheme leads to straightforward computation of marginal likelihoods based on the candidate's formula [3] .

Acknowledgments

We are grateful to Rudi Frühwirth for providing the normal mixture approximation to the log exponential distribution which is in fact an essential ingredient for our Gibbs sampling scheme. Furthermore we want to thank Stefan Höglinger from the KfV, the Austrian Road Safety Board, for providing the data analyzed in Subsection 4.

References

- [1] Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* 6, 251–262.
- [2] Albert, J. H. (1992). A Bayesian analysis of a Poisson random-effects model. *American Statistician* 46, 246–253.
- [3] Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90, 1313–1321.
- [4] Chib, S., E. Greenberg, and R. Winkelmann (1998). Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics* 86, 33–54.
- [5] Chib, S., F. Nardari, and N. Shephard (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics* 108, 281–316.
- [6] Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika* 70, 269–274.
- [7] Frühwirth-Schnatter, S. and H. Wagner (2005). Auxiliary Mixture Sampling for Parameter-driven Models of Time Series of Counts with Applications to State Space Modelling. Research Report IFAS 2005-11, <http://www.ifas.jku.at>.
- [8] Gelfand, A., S. Sahu, and B. Carlin (1995). Efficient parametrisations for normal linear mixed models. *Biometrika* 82, 479–488.
- [9] Hurn, M., A. Justel, and C. P. Robert (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics* 12, 55–79.
- [10] Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies* 65, 361–393.

- [11] Lenk, P. J. and W. S. DeSarbo (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika* 65, 93–119.
- [12] McCullagh, P. and J. A. Nelder (1999). *Generalized linear models*. Chapman & Hall Ltd.
- [13] Robert, C. P. and G. Casella (1999). *Monte Carlo statistical methods*. Springer Series in Statistics. New York/Berlin/Heidelberg: Springer-Verlag Inc.
- [14] Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika* 81, 115–131.
- [15] Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82, 528–540.
- [16] van Dyk, D. and X.-L. Meng (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics* 10, 1–50.
- [17] Wang, P., M. L. Puterman, I. Cockburn, and N. Le (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics* 52, 381–400.
- [18] Zeger, S. and M. Karim (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* 86, 79–86.
- [19] Zeger, S. L. and B. Qaqish (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics* 44, 1019–1031.
- [20] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.

L_p -norm Estimation: Some Simulation Studies in Presence of Multicollinearity

Massimiliano Giacalone¹ and Rino Richiusa

¹ *University of Naples "Federico II", Naples, Italy*
Email: mgiacalo@unina.it

Abstract : In this paper we propose some simulation studies in order to compare the L_p -norm estimators with the Least Squares method (L_2), introducing a linear regression model in the case of multicollinearity. A comparative analysis is made applying three estimation methods to evaluate the empirical distribution of the regression parameters. Looking at the simulation results we note that improvement using L_p estimators instead of L_2 is more evident in the case of medium collinearity than in the case of strong one.

Key words: Multicollinearity, exponential power function, L_p norm estimators, kurtosis indexes.

1 Introduction

“For non-normal distributions of the ϵ_i ’s the least-squares estimates have minimum variance among unbiased estimates that are linear combinations of the Y_i ’s” (Cox and Hinkley, 1968). And in case of multicollinearity?

In this paper we consider the possibility of creating some adaptive robust estimation procedures for the standard linear regression model when the disturbance vector has deviated from normality.

We propose some simulation planes in order to compare the L_p -norm estimators with the least squares method in the case of multicollinearity, a common regression problem introduced in the second section of the paper.

After a brief review on the origins of the Exponential Power Function (E.P.F.), a useful family of symmetrical random error curves, we show their connection with the L_p -norm methods.

In the fourth section we investigate the relationships between the different values of p , the shape parameter of the E.P.F., and some kurtosis indexes that refer to the error distributions, then we analyze and propose the $L_{p_{\min}}$ algorithm.

In the subsequent sections we compare three estimation procedures by introducing a linear regression model with collinear regressors. The aim of the simulation study is to choose the best rule for selecting the most appropriate value of p for any given error distribution and to evaluate the influence of multicollinearity in the parameter estimation procedures in terms of bias and variance. Finally, in the last paragraph, some results and comparisons are discussed.

2 Multicollinearity in regression

A regression model has two main purposes: 1) it could be used to find out to what extent the outcome (dependent variable) can be predicted by the independent variables and 2) it could be used to determine the strength of a theoretical relationship between the dependent variable and the independent variables.

A regression problem can be affected by multi-collinearity. The absence of multi-collinearity is essential to obtain optimal estimates by a multiple regression model. In regression the multi-collinearity or collinearity problem occurs when several regressors are highly correlated. When things are related, we say they are linearly dependent on each other because one can nicely fit a straight regression line to pass through many of the data points of those variables. Collinearity simply means co-dependence. It would have to be eliminated. Doing it is problematic when one's purpose is explanation rather than mere prediction. Indeed when prediction is the goal no problem arises if, among dependent variables, two regressors have the same 'meaning': simply it is possible eliminating one. On the contrary, when explanation is the goal regressors have been selected according a theoretical rule, so that it is possible eliminate no variable because of the model is theoretical, that is given; it has to be explained exactly, including all selected variables.

Therefore the multi-collinearity problem can and must be eliminated in the case of prediction while is often uneliminable in the case of explanation. In this latter case eliminating a variable leads to a problem of bad-specification of the model. A viable remedy for bad-specification is to prefer a different estimator than the least squares one. This is because the 'classical' Ordinary Least Squares (OLS) method provides estimates which could be statistically not significant in presence of multi-collinearity. It would be better to use alternative methods such as the Ridge Regression (Morris, 1982; Pagel and Lunneberg, 1985) or the Partial Least Squares (Cassel et al., 1999). Here we are going to carry out some simulation studies in order to show some evident differences among three estimation methods (Least Squares estimators and two different L_p -norm estimators) in presence of multi-collinearity.

3 The Exponential Power Function and the L_p -norm estimators

The E.P.F is a family of probability functions proposed by Subbotin in 1923 and studied by Vianelli (1963), Lunetta (1963), Mineo (1989). The density function is:

$$f_p(z) = \frac{1}{2p^{1/p}\sigma_p\Gamma(1+1/p)} \exp \left[-\frac{1}{p} \left| \frac{z - M_p}{\sigma_p} \right|^p \right] \quad (1)$$

where $M_p = E(z)$ is the *location parameter*, $\sigma_p = (E[|z - M_p|])^{1/p}$ is the *scale parameter*, and $p > 0$ is the *shape parameter*.

Considering the Pearson kurtosis index β_2 we distinguish:

- $0 < p < 1$: double exponential distributions, $\beta_2 > 6$;
- $1 < p < 2$: leptokurtic distributions, $3 < \beta_2 < 6$;
- $p > 2$: platikurtic distributions, $1.8 < \beta_2 < 3$;

For particular values of p we have: the *Laplace distribution* ($p = 1$, $\beta_2 = 6$); the *Gaussian distribution*, ($p = 2$, $\beta_2 = 3$); and the *Uniform distribution* ($p \rightarrow \infty$, $\beta_2 \rightarrow 1.8$).

Let us consider a sample of n observed data (y_i, x_i) , a general linear regression model is:

$$y_i = g(x_i, \theta) + \epsilon_i, \quad (2)$$

with g a linear function.

The L_p -norm estimators are a mere generalization of the Least Squares replacing exponent 2 by a general exponent p . Therefore they minimize the sum of the p -th power of the absolute deviations of the observed points from the regression function:

$$S_p(\theta) = \sum_{i=1}^n |y_i - g(x_i, \theta)|^p \quad 1 \leq p < \infty \quad (3)$$

Under the regularity assumptions the log-likelihood related to the sample is given by:

$$l(\theta, \sigma_p, p) = -n \log [2p^{1/p}\sigma_p\Gamma(1+1/p)] - [(p\sigma_p)^{-1} \sum |y_i - g(x_i, \theta)|^p] \quad (4)$$

where we consider $z = y_i$ and $\mu_p = g(x_i, \theta)$

$$\begin{aligned} \delta l / \delta \theta_j &= \sum_{i=1}^n |y_i - g(x_i, \theta)|^{p-1} \text{sign}(y_i - g(x_i, \theta)) \frac{\delta g}{\delta \theta_j} \\ \sum_{i=1}^n |y_i - g(x_i, \theta)|^p &= \min \quad \text{with } p \geq 1 \end{aligned} \quad (5)$$

The optimal exponent p for the L_p -norm estimators of the regression parameters is the shape parameter p of the E.P.F.

If it is unknown we have two related problems to consider:

- (1) The estimation of the *exponent* p on the sample data.

- (2) The choice of the minimization algorithm to obtain the *regression parameters* estimation.

To estimate the exponent p it is possible to find out the following proposals:

- Harter (1977), noting that p depends on $\hat{\beta}_2$ (the sample residual Kurtosis), proposed selecting p with the following rule:
 if $\hat{\beta}_2 > 3.8$ use $p = 1$ (the *least absolute deviations* regression).
 if $2.2 < \hat{\beta}_2 < 3.8$ use $p = 2$ (the *least squares* regression).
 if $\hat{\beta}_2 < 2.2$ use $p = \infty$ (the *minimax* or *Chebychev* regression).
- Money et al. (1982) and Sposito et al. (1983) proposed two different criteria respectively:

$$\hat{p} = 9/\hat{\beta}_2^2 + 1 \quad \text{for } 1 \leq p < \infty \quad (6)$$

$$\hat{p} = 6/\hat{\beta}_2 \quad \text{for } 1 < p < 2 \quad (7)$$

- Mineo A. (1989) proposed the Generalized Kurtosis β_K , as described below.
- Mineo A.M. (1994) considered a new method to estimate p , based on an empirical index called VI.
- Agrò (1995) proposed a maximum likelihood estimation either for the regression parameters or for the p shape parameter.

4 The Exponential Power Function kurtosis indexes

For the density (1), the theoretical moment of order k is a function of the shape parameter p as follows:

$$E|z - M_p|^k = (p\sigma_p^2)^{-k/p} \frac{\Gamma((k+1)/p)}{\Gamma(1/p)} = \mu_k \quad (8)$$

The *ratios of the moments of order $2k$ and the squared moment of order k* only depend on the shape parameter p . This theoretic relation is also called "*Generalized Kurtosis*" (Mineo, 1989) :

$$\beta_k = \frac{\mu_{2k}}{\mu_k^2} = \frac{\Gamma(1/p) \Gamma((2k+1)/p)}{[\Gamma((k+1)/p)]^2}$$

if $k = 2$ we can write the *Pearson Kurtosis* index:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\Gamma(1/p) \Gamma(5/p)}{[\Gamma(3/p)]^2} \quad (9)$$

If $k = 1$ considering the square root of the reciprocal we get the *Geary length of tails* index:

$$I = \frac{\mu_1}{\sqrt{\mu_2}} = \frac{\Gamma(2/p)}{\sqrt{\Gamma(1/p) \Gamma(3/p)}} \quad (10)$$

The indexes I and β_2 show a different behaviour according to the variation of p (Giacalone, 1997).

Calculating the sample values of I and β_2 , it is possible to obtain, by inverse interpolation, two different estimations of p .

Gonin and Money (1987), Lunetta, (1966), Kendall-Stuart, (1966) considered the unbiased estimates of the second and fourth order sample moments with correction factors depending on the sample size n :

$$\hat{\mu}_2 = \frac{1}{n-1} \sum_i (\epsilon_i - \bar{\epsilon})^2$$

$$\hat{\mu}_4 = \frac{(n^2 - 2n + 3)}{(n-1)(n-2)(n-3)} \sum_i (\epsilon_i - \bar{\epsilon})^2 - \frac{3(n-1)(2n-3)}{n(n-2)(n-3)} \hat{\mu}_2^2$$

The ratio of $\hat{\mu}_4$ and $\hat{\mu}_2$ gives the following estimator of β_2 :

$$\hat{\beta}_2 = \frac{\hat{\mu}_4}{\hat{\mu}_2^2} \quad (11)$$

For the I empirical index we obtain:

$$\hat{I} = \frac{\sum_i |\epsilon_i - \bar{\epsilon}|}{\sqrt{\sum_i |\epsilon_i - \bar{\epsilon}|^2}} \frac{\sqrt{n-1}}{n} \quad (12)$$

5 The $L_{p_{\min}}$ algorithm

It is based on a two-steps alternating procedure : a) minimization procedure to estimate the parameters, b) joint inverse function of I and β_2 to estimate p .

The algorithm is stopped when p does not vary significantly.

The function used to estimate p is therefore the following:

$$\left[(I - \hat{I}) : 0.86054 \right]^2 + \left[(\beta_2 - \hat{\beta}_2) : 25.2 \right]^2 = \min, \quad (13)$$

where I , $\hat{I}$, β_2 , $\hat{\beta}_2$ are respectively given by (10), (12), (9), (11). For simplicity we express the (13) as $[f(p)]^2 + [g(p)]^2 = \min$ to take into account the different variability and average order size related to p , and the standardization factors are the maximum theoretical values.

So using the relation (9) we calculate $\max(\beta_2) = 25.2$ for $p = 0.5$, lower bound in our simulation plan, whilst using the relation (10) we calculate $\max(I) = 0.86054$, for $p = 10$, upper bound in our simulation plan.

The proposed algorithm (Giacalone, 1997) is then specified in the following steps:

- **STEP 0:** Set $i = 0$ and $p_0 = 2$;
- **STEP 1:** Fit the model to the data using the previous step value p_i ;
- **STEP 2:** Compute the estimated residuals $\epsilon_i = y_i - g(x_i, \theta)$, their average $\bar{\epsilon}$ and insert these quantities in the (13) which is equal to the sum of the two squared

functions to be minimized;

- **STEP 3:** Minimize the function (13) to obtain p_{i+1} , new estimate of p ;
- **STEP 4:** Compare the estimated p_{i+1} with the previous p_i , and if $|p_{i+1} - p_i| > 0.01$ then set $i = i + 1$ and repeat steps 1-4, otherwise:
- **STEP 5:** Stop the algorithm assuming the values $\hat{\theta}_i = \theta_{ij}$ as L_p -norm estimators for the parameters θ_i and the value $p = p_i$ as joint estimation of p .

In step 1 a nonlinear L_p -norm estimation is considered (Fletcher and Reeves, 1964). In step 3 a parabolic interpolation method (Everitt, 1987) to find the minimum of the sum of squared functions (13) is adopted.

6 The simulation plan

We consider 500 samples of sizes $n = 50, 100$, generated from E.P.F., and 6 values of p , ranging from 1.0 to 3.5 with step 0.5. The algorithm for generating the ϵ_i (for $p \geq 1$) from an E.P.F. is suggested by Chiodi (1986).

The values of y_i are given by the multiple regression model:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \quad (14)$$

with X_1, X_2 identically distributed and independent variables from a Gaussian standardized distribution and X_3 a linear combination of X_1 and X_2 .

$$X_3 = X_1 + X_2 + Z \text{ with } Z \sim N(0, \sigma_z^2) \quad (15)$$

Therefore we can write the related variance and covariance matrix: It is easy to see

$$\begin{vmatrix} & X_1 & X_2 & X_3 \\ X_1 & 1 & 0 & 1 \\ X_2 & 0 & 1 & 1 \\ X_3 & 1 & 1 & 2 + \sigma_z^2 \end{vmatrix}$$

that: $E(X_3^2) = E(X_1^2) + E(X_2^2) + \sigma_z^2 = 2 + \sigma_z^2$

and the correspondent correlation matrix is equal to:

$$r = \begin{vmatrix} 1 & 0 & 1/\sqrt{2 + (\sigma_z^2)} \\ 0 & 1 & 1/\sqrt{2 + (\sigma_z^2)} \\ 1/\sqrt{2 + (\sigma_z^2)} & 1/\sqrt{2 + (\sigma_z^2)} & 1 \end{vmatrix}$$

where $r_{13} = \text{cov}(X_1, X_3)/\sqrt{\text{var}(X_1)\text{var}(X_3)} = 1/\sqrt{1(2 + \sigma_z^2)} = r_{23}$

and $R_{3.12}^2 = 1 - \det A / \det A_{33} = 2/(2 + \sigma_z^2)$

In the simulation model one can see that the rate of multicollinearity is (inversely)

proportional to σ_z^2 .

Let be the parameter values $\beta_0 = 1, \beta_1 = 2, \beta_2 = 3, \beta_3 = 4$, a comparative analysis is made applying the following three estimation methods on the same samples, with the results reported in the Tab 2-7 :

- (1) Least squares estimators (L_2).
- (2) L_p -norm estimators with theoretical p of the E.P.F. (L_p).
- (3) L_p -norm estimators with p as in our proposal (Section 5) ($L_{p_{\min}}$).

One can see that for any p and for any method, the parameter estimates of $\beta_0, \beta_1, \beta_2, \beta_3$, are biased when $n = 50$ but their variances decrease for increasing values of n . That depends on the multicollinearity of the model. It is possible noting the unbiasedness of the estimates only for middle-large samples sizes.

The L_p -norm estimators give us better parameter estimates for the parameters β_1 and β_2 compared to the least squares method especially for p far from 2 (see tab 7 and tab 8).

The gain in efficiency using the L_p -norm estimators is higher for the L_p method in all the cases considered except for the case $p = 2$ where the error is generated by a Gaussian distribution.

The $L_{p_{\min}}$ method could be considered as half-way between the L_2 and L_p methods because we estimate the exponent p on the sample data.

CASE 1			$\sigma_z = 1$			$R_{3,12}^2 = 0.66$		
p	$M(\beta_0)$	$V(\beta_0)$	$M(\beta_1)$	$V(\beta_1)$	$M(\beta_2)$	$V(\beta_2)$	$M(\beta_3)$	$V(\beta_3)$
$n = 50$								
1.0	1.162	.928	2.117	.964	3.124	.849	3.748	1.334
1.5	0.973	.816	1.877	.834	2.889	.826	4.042	1.083
2.0	1.029	.799	1.956	.786	2.835	.778	3.987	1.037
2.5	1.037	.678	1.934	.713	2.903	.645	4.059	0.968
3.0	0.956	.665	1.873	.685	2.918	.616	3.764	0.864
3.5	1.235	.558	2.097	.609	3.221	.593	3.952	0.798
$n = 100$								
1.0	1.043	.128	1.957	.164	2.934	.024	3.838	.033
1.5	1.034	.116	1.942	.153	2.919	.022	4.039	.028
2.0	1.018	.099	2.001	.126	2.965	.017	4.022	.023
2.5	1.025	.078	2.034	.101	3.015	.014	3.959	.019
3.0	1.028	.065	1.977	.078	2.953	.011	3.853	.016
3.5	1.039	.058	1.963	.069	2.941	.009	3.978	.014

Table 1: Mean and Variance of $\beta_0, \beta_1, \beta_2, \beta_3$ for a multiple regression model ($\beta_0 = 1, \beta_1 = 2, \beta_2 = 3, \beta_3 = 4$) estimated with L_p -method on 500 samples of size $n = 50, n = 100$.

CASE 2			$\sigma_z = 0.01$				$R_{3,12}^2 = 0.99$	
p	$M(\beta_0)$	$V(\beta_0)$	$M(\beta_1)$	$V(\beta_1)$	$M(\beta_2)$	$V(\beta_2)$	$M(\beta_3)$	$V(\beta_3)$
$n = 50$								
1.0	1.276	1.281	2.202	1.323	3.135	1.481	3.801	1.634
1.5	1.053	1.016	1.856	1.053	2.859	1.213	4.052	1.466
2.0	1.049	.949	1.942	1.002	2.845	1.187	3.963	1.255
2.5	1.047	.878	1.921	.938	2.883	1.045	4.072	1.109
3.0	0.934	.723	2.143	.801	3.108	.912	3.792	1.021
3.5	1.262	.622	2.104	.712	3.221	.796	3.938	0.934
$n = 100$								
1.0	1.067	.625	1.944	.744	2.944	.834	3.855	.941
1.5	1.045	.514	1.931	.633	2.931	.711	4.043	.844
2.0	1.022	.397	2.005	.503	2.922	.566	4.051	.654
2.5	1.037	.255	2.051	.401	3.029	.533	4.065	.592
3.0	1.045	.155	1.983	.277	2.934	.345	3.839	.416
3.5	1.061	.103	1.942	.188	2.928	.233	4.044	.307

Table 2: Mean and Variance of $\beta_0, \beta_1, \beta_2, \beta_3$ for a multiple regression model ($\beta_0 = 1, \beta_1 = 2, \beta_2 = 3, \beta_3 = 4$) estimated with L_p -method on 500 samples of size $n = 50, n = 100$.

CASE 3			$\sigma_z = 1$				$R_{3,12}^2 = 0.66$	
p	$M(\beta_0)$	$V(\beta_0)$	$M(\beta_1)$	$V(\beta_1)$	$M(\beta_2)$	$V(\beta_2)$	$M(\beta_3)$	$V(\beta_3)$
$n = 50$								
1.0	1.143	.865	2.111	.883	3.133	.821	3.802	1.277
1.5	0.981	.802	1.854	.799	2.898	.766	4.032	1.055
2.0	1.033	.745	1.949	.746	2.866	.722	3.985	0.965
2.5	0.995	.655	1.912	.695	2.915	.601	4.044	0.929
3.0	0.967	.624	1.892	.633	2.932	.587	3.811	0.833
3.5	1.195	.507	2.075	.576	3.198	.544	3.898	0.776
$n = 100$								
1.0	1.028	.101	1.968	.134	2.955	.021	3.854	.031
1.5	1.025	.099	1.955	.124	2.901	.019	3.998	.026
2.0	1.013	.093	1.998	.119	2.977	.015	4.018	.021
2.5	1.019	.075	2.029	.098	3.011	.012	3.966	.017
3.0	1.024	.062	1.975	.074	2.951	.009	3.866	.013
3.5	1.033	.056	1.961	.067	2.939	.007	3.975	.011

Table 3: Mean and Variance of $\beta_0, \beta_1, \beta_2, \beta_3$ for a multiple regression model ($\beta_0 = 1, \beta_1 = 2, \beta_2 = 3, \beta_3 = 4$) estimated with L_p -method on 500 samples of size $n = 50, n = 100$.

CASE 4		$\sigma_z = 0.01$				$R_{3,12}^2 = 0.99$		
p	$M(\beta_0)$	$V(\beta_0)$	$M(\beta_1)$	$V(\beta_1)$	$M(\beta_2)$	$V(\beta_2)$	$M(\beta_3)$	$V(\beta_3)$
$n = 50$								
1.0	1.164	.921	2.122	.945	3.119	.841	3.744	1.329
1.5	0.969	.820	1.875	.831	2.878	.822	4.039	1.077
2.0	1.025	.794	1.958	.783	2.832	.769	3.978	1.032
2.5	1.035	.677	1.936	.709	2.915	.643	4.044	0.958
3.0	0.944	.655	1.876	.679	2.921	.620	3.788	0.871
3.5	1.244	.566	2.098	.611	3.219	.588	3.961	0.787
$n = 100$								
1.0	1.045	.122	1.944	.161	2.941	.023	3.855	.032
1.5	1.031	.115	1.941	.149	2.933	.023	4.041	.027
2.0	1.009	.095	1.999	.121	2.971	.016	4.019	.021
2.5	1.024	.076	2.031	.099	3.018	.013	3.962	.018
3.0	1.025	.062	1.981	.074	2.951	.011	3.844	.015
3.5	1.034	.056	1.973	.067	2.945	.008	3.981	.013

Table 4: Mean and Variance of $\beta_0, \beta_1, \beta_2, \beta_3$ for a multiple regression model ($\beta_0 = 1, \beta_1 = 2, \beta_2 = 3, \beta_3 = 4$) estimated with L_p -method on 500 samples of size $n = 50, n = 100$.

CASE 5		$\sigma_z = 1$				$R_{3,12}^2 = 0.66$		
p	$M(\beta_0)$	$V(\beta_0)$	$M(\beta_1)$	$V(\beta_1)$	$M(\beta_2)$	$V(\beta_2)$	$M(\beta_3)$	$V(\beta_3)$
$n = 50$								
1.0	1.155	1.098	2.122	1.088	3.136	1.023	3.698	1.529
1.5	0.955	.905	1.855	.923	2.901	.958	4.051	1.208
2.0	1.033	.745	1.949	.746	2.866	.722	3.985	0.965
2.5	1.031	.701	1.921	.751	2.922	.688	4.123	0.955
3.0	0.942	.688	1.855	.713	2.944	.651	3.728	0.888
3.5	1.303	.599	2.102	.654	3.234	.611	3.911	0.833
$n = 100$								
1.0	1.041	.256	1.932	.231	2.569	.058	3.855	.053
1.5	1.031	.178	1.929	.189	2.933	.039	4.055	.037
2.0	1.013	.093	1.998	.119	2.977	.015	4.018	.021
2.5	1.033	.069	2.043	.109	3.022	.013	3.961	.020
3.0	1.033	.066	1.966	.081	2.966	.012	3.899	.017
3.5	1.044	.061	1.944	.072	2.955	.011	3.985	.015

Table 5: Mean and Variance of $\beta_0, \beta_1, \beta_2, \beta_3$ for a multiple regression model ($\beta_0 = 1, \beta_1 = 2, \beta_2 = 3, \beta_3 = 4$) estimated with L_p -method on 500 samples of size $n = 50, n = 100$.

CASE 6			$\sigma_z = 0.01$			$R_{3,12}^2 = 0.99$		
p	$M(\beta_0)$	$V(\beta_0)$	$M(\beta_1)$	$V(\beta_1)$	$M(\beta_2)$	$V(\beta_2)$	$M(\beta_3)$	$V(\beta_3)$
$n = 50$								
1.0	1.164	.921	2.121	.967	3.134	.855	3.337	1.345
1.5	0.988	.820	1.888	.836	2.885	.829	4.041	1.091
2.0	1.025	.794	1.958	.783	2.832	.769	3.978	1.032
2.5	1.039	.733	1.942	.737	2.921	.691	4.077	0.991
3.0	0.966	.703	1.887	.701	2.933	.612	3.755	0.903
3.5	1.246	.588	2.115	.651	3.233	.609	3.945	0.847
$n = 100$								
1.0	1.044	.145	1.966	.188	2.945	.029	3.844	.038
1.5	1.036	.121	1.951	.169	2.901	.024	4.045	.032
2.0	1.009	.095	1.999	.121	2.971	.016	4.019	.021
2.5	1.022	.082	2.039	.105	3.021	.015	3.933	.020
3.0	1.027	.073	1.975	.083	2.977	.013	3.799	.017
3.5	1.033	.062	1.961	.071	2.933	.011	3.965	.015

Table 6: Mean and Variance of $\beta_0, \beta_1, \beta_2, \beta_3$ for a multiple regression model ($\beta_0 = 1, \beta_1 = 2, \beta_2 = 3, \beta_3 = 4$) estimated with L_p -method on 500 samples of size $n = 50, n = 100$.

CASE 1,3,5				$\sigma_z = 1$				$R_{3,12}^2 = 0.66$				
p	1.0		1.5		2.0		2.5		3.0		3.5	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
$n = 50$												
L_p	2.2	1.8	1.3	1.2	1.0	1.0	1.1	1.2	1.4	1.2	1.8	1.3
$L_{p_{\min}}$	2.1	1.7	1.3	1.4	0.8	0.9	0.8	0.9	1.3	1.2	1.7	1.3
$n = 100$												
L_p	1.8	1.6	1.4	1.1	1.0	1.0	1.1	1.1	1.2	1.1	1.4	1.5
$L_{p_{\min}}$	1.8	1.5	1.3	1.0	0.9	0.9	1.1	1.0	1.2	1.1	1.3	1.3

Table 7: Relative efficiency of L_p -norm estimators compared to the least squares (parameter β_1, β_2).

CASE 2,4,6					$\sigma_z = 0.01$				$R_{3.12}^2 = 0.99$			
p	1.0		1.5		2.0		2.5		3.0		3.5	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
$n = 50$												
L_p	1.2	1.1	1.2	1.2	1.0	1.0	1.1	1.2	1.3	1.2	1.5	1.2
$L_{p_{\min}}$	1.3	1.2	1.3	1.3	0.9	0.8	0.9	1.1	1.4	1.2	1.4	1.2
$n = 100$												
L_p	1.2	1.1	1.1	1.1	1.0	1.0	1.2	1.1	1.2	1.1	1.3	1.1
$L_{p_{\min}}$	1.1	1.1	1.2	1.2	1.1	0.9	1.0	1.0	1.2	1.1	1.3	1.2

Table 8: Relative efficiency of L_p -norm estimators with respect to the least squares (parameter β_1, β_2).

7 Conclusions

In this ending section we underline the objectives reached with our simulation plan. In this study the L_p -norm methods are considered not in order to reduce the multicollinearity in the model, but with the aim of seeing the improvements in the parameter estimation.

Looking at the simulation results we note that the improvement using L_p estimators instead of least squares is more evident in the case $\sigma_z = 1$ $R_{3,12}^2 = 0,66$ (medium multicollinearity) than in the case $\sigma_z = 0,01$ $R_{3,12}^2 = 0,99$ (strong multicollinearity).

That is due to the characteristics of L_p methods that are adaptive procedures related to the erratic component of the model and not related to the deterministic one.

Finally, the simulation studies show how our algorithm ($L_{p_{\min}}$) achieves more efficient estimates for the regression parameters when compared with the least squares procedure.

In particular, using the L_p and the $L_{p_{\min}}$ method, a better performance, in terms of the variances of the parameter estimates, is always obtained in the case of nonnormal symmetric distributions compared the least squares situation also considering a model with collinear regressors.

Acknowledgments

This paper was partially financed by the P.R.I.N 2000 Project *Data Mining and Symbolical Data Analysis* on M.U.R.S.T. funds; National Coordinator: Prof. N. Carlo Lauro, University of Naples "Federico II".

References

- [1] Agrò G. (1995): "Maximum likelihood estimation for the exponential power function parameters". *Comm. Statist - Simula. Computa.*, 24, pp. 523,536.
- [2] Cassel, C.; Westlund, A. H. and Hackl, P. (1999): "Robustness of partial least-

- squares method for estimating latent variable quality structures". *Journal of Applied Statistics*, 26, 435-448.
- [3] Chiodi M. (1986): "Procedures for Generating Pseudo-Random Numbers from a Normal Distribution of Order p " *Riv. Stat. Applicata*, 19, pp. 7-26.
 - [4] Cox D.R., Hinkley D.V. (1968): "A note on the Efficiency of Least Squares Estimates". *Journal of Royal Statistical Society, B*; 30, 841.
 - [5] Everitt B. and Hand D. (1987): "The Statistical Consultant in Action", New York: Cambridge University Press.
 - [6] Fletcher R. and Reeves C.M. (1964): "Function minimization by conjugate gradients" *Comput. J.*, 7, pp. 149-154.
 - [7] Giacalone M. (1997): " L_p -norm estimators for nonlinear regression models" *Student*, vol 2, pp 119-130.
 - [8] Gonin R. and Money A.H. (1987): "A Review of Computational Methods for Solving the Nonlinear L_1 Norm Estimation Problem". *S.D.A.B. on the L_1 Norm and Related Methods*. Dodge ed al., North Holland.
 - [9] Harter H.L. (1977): "The nonuniqueness of absolute values regression". *Commun. Statist. Simula. Computa.* A6, pp. 829-838.
 - [10] Kendall M.C. and Stuart A. (1966): "The Advanced Theory of Statistics" *Griffin*, London.
 - [11] Lunetta G. (1963): "Di una Generalizzazione dello Schema della Curva Normale" *Ann. Facoltà Econ. Comm.* Palermo, 17, pp. 237-244.
 - [12] Lunetta G. (1966): "Di Alcune Distribuzioni Deducibili da una Generalizzazione dello Schema della Curva Normale". *Ann. Facoltà Econ. Comm.* Palermo, 20, pp. 117-143.
 - [13] Mineo A. (1989): "The Norm- p Estimation of Location, Scale and Simple Linear Regression Parameters". *Lectures Notes in Statistics-Statistical Modelling* Trento Pringer-Verlag, pp. 222-223.
 - [14] Mineo A.M. (1994): "Un nuovo metodo di stima di p per una corretta valutazione dei parametri di intensità e di scala di una curva normale di ordine p ". *Atti della XXXVII Riunione Scientifica della SIS*, vol.2, San Remo pp. 147-154.
 - [15] Money A.H. et al. (1982): "The Linear Regression Model: L_p -Norm Estimation and the Choice of p ". *Commun. Statist. - Simula. Computa*, 11, pp. 89-109.
 - [16] Morris, J.D. (1982): "Ridge regression and some alternative weighting techniques: A comment on Darlington." *Psychological Bulletin*, 91, 203-210.
 - [17] Pagel, M.D. and Lunneberg, C.E. (1985): "Empirical evaluation of ridge regression". *Psychological Bulletin*, 97, 342-355.
 - [18] Sposito V.A. (1983): "On Unbiased L_p Regression". *J.A.S.A.*, 77.
 - [19] Subbotin M. (1923): "On the Law of Frequency of Error". *Mathematicheskii Sbornik*.
 - [20] Vianelli S. (1963): "La Misura della Variabilità Condizionata in uno Schema Generale delle Curve Normali di Frequenza". *Statistica*, XXIII n. 4.

Assessing a Vector Parameter

Donald A. S. Fraser and Nancy Reid

*Department of Statistics, University of Toronto
100 St. George Street, Toronto, Canada M5S 3G3*

E-mail: dfraser@utstat.toronto.edu; reid@utstat.utoronto.ca

Abstract: We describe a conditional p -value approximation for inference about a vector parameter θ . This measures the magnitude of the departure from an hypothesized value θ_0 , conditional on the direction of departure. The form of the approximation is similar to the Lugannani and Rice approximation to p -values for scalar parameters, and closely related to Skovgaard's (2001) version of w^* .

Key words: Ancillary; Conditional likelihood; Composite hypothesis; Conditioning; Conical test; Directional test; Vector interest parameter; Likelihood; Marginal likelihood; Profile likelihood; p -value.

1 Introduction

The assessment of a vector parameter is central to statistical theory. The analysis of variance with tests and confidence regions for treatment effects is well established and the related distribution theory is conveniently quite straightforward, particularly in the normal error case. In more general contexts such as generalized linear models, the assessment is usually based on the likelihood ratio or maximum likelihood departure measures and the related distribution theory is first order. As such, and confirmed by experience, both the type of departure and the distributional accuracy can be strongly misleading.

Some improvements in the departure measure and the distributional accuracy have been obtained with directional tests (Fraser & Massam, 1985; Skovgaard, 1988; Cheah *et al.*, 1994) but these typically require an exponential model.

For greater generality we now examine a continuous statistical model with parameter θ of dimension p and interest parameter $\psi(\theta)$ of dimension d . We examine an observed data point y^0 recorded on a score variable space which has suitable linearity, and we consider the directed departure of the data point from expectation under a

hypothesized value ψ for $\psi(\theta)$. The directed p -value is obtained as the observed conditional distribution function for the distance of the score variable from its expected value, given the direction of the departure. We view this as arguably the appropriate statistical assessment of the data point with respect to the parameter value ψ ; for some background see Fraser & Massam (1985).

This distribution function value can be approximated by

$$p(\psi) = G_d(\chi^0) + b_d \{G_{d+1}(\chi^0) - G_d(\chi^0)\} \delta(\chi^0) \quad (1.1)$$

where $G_d(\chi)$ is the ordinary chi distribution function

$$G_d(\chi) = \int_0^\chi c_d t^{d-1} e^{-t^2/2} dt \quad (1.2)$$

with norming constant $c_d = 1/\Gamma(d/2)2^{d/2-1}$, and χ^0 is the observed value obtained from the score-based departure measure described in §3. The constant $b_d = c_d/c_{d+1} = 2^{1/2}\Gamma\{(d+1)/2\}/\Gamma(d/2)$ is the ratio of norming constants from adjacent χ density functions. The adjustment factor $\delta(\chi)$, obtained from (3.4) and (3.5), provides a measure of how much the underlying distribution differs from normality.

The preceding p -value has second order accuracy, by which we mean the error of the approximation to the exact distribution is $O(n^{-1})$. The usual chi-squared approximations have relative error $O(n^{-1/2})$. The approximation is derived in §3 after a brief summary of needed likelihood results in §2; it has second order accuracy and has components that are accessible and easily interpreted. By comparison, the Skovgaard (1988) directional test is also accurate to second order but requires full exponential form and cumulant type calculations, and the Cheah *et al.* (1994) directional test is third order but also requires full exponential model form and more detailed model characteristics. The proposed test combines the likelihood ratio and score test statistics in a manner not unrelated to that found with the Lugannani & Rice (1980) approximation formula. Examples are discussed in §4.

The Bartlett (1955) correction of the likelihood ratio test can also be used for assessing vector parameters. This has the attractive property that the distribution ascribed to the related pivot has $O(n^{-2})$ distributional accuracy and the procedure can then be described as accurate to fourth order. Our viewpoint is that the assessment should take account of where the data point is relative to a parameter value, in much the same way as recent third order p -values do for a scalar parameter of interest, by recording the percentile position of the data point relative to this parameter. From this viewpoint we argue in §5 that the Bartlett corrected likelihood ratio assessment is misleading if viewed as being of other than first order. Skovgaard (2001) derives an adjustment to the likelihood ratio statistic somewhat analogous to that derived here, although from a different point of view. The relationship between the two approaches deserves further study.

2 Likelihood Background

Suppose we have a statistical model with density $f(y; \theta)$, where $y = (y_1, \dots, y_n)'$ and θ is p -dimensional. Recent third order theory for assessing scalar or vector parameters can be based on just two ingredients, $\ell(\theta)$ and $\varphi(\theta)$, where $\ell(\theta) = \log f(y^0; \theta)$ is the observed log-likelihood function and $\varphi(\theta) = (d/dV)\ell(\theta; y)|_{y^0}$ is the observed log-likelihood gradient, calculated in directions $(v_1, \dots, v_p) = V$ tangent to an exact or approximate ancillary. The parameter $\varphi(\theta)$ is for convenience written as a row vector, and more explicitly we have

$$\begin{aligned}\varphi(\theta) &= \left\{ \frac{d}{dv_1} \ell(\theta; y^0), \dots, \frac{d}{dv_p} \ell(\theta; y^0) \right\} \\ &= \left\{ \frac{d}{dt} \ell(\theta; y^0 + tv_1) \Big|_{t=0}, \dots, \frac{d}{dt} \ell(\theta; y^0 + tv_p) \Big|_{t=0} \right\}\end{aligned}$$

where the $n \times 1$ vectors v_j make up the columns of the $n \times p$ matrix V .

With an appropriate n -dimensional pivotal quantity $z(y, \theta)$ we can obtain V from the expression

$$V = -z_{y'}^{-1}(y^0, \hat{\theta}^0) z_{\theta'}(y^0, \hat{\theta}^0),$$

where the subscripts denote partial differentiation with respect to the row vectors y' and θ' . The use of the gradient implements conditioning on an approximate ancillary statistic (Fraser & Reid, 2001), and several examples of pivotal quantities z are illustrated in Fraser, Reid & Wu (1999), Fraser, Wong & Wu (1999) and Reid (2003).

Third order p -values for a scalar parameter $\psi(\theta)$ can be obtained from just $\{\ell(\theta), \varphi(\theta)\}$ using the likelihood ratio r and a special maximum likelihood departure q ; see, for example, equations (1.5) and (2.11) in Fraser, Reid & Wu, or equations (3.13) and (3.16) in Reid (2003).

The two ingredients $\ell(\theta)$ and $\varphi(\theta)$ also yield an approximation to the original model that produces equivalently the third order p -values and likelihoods just described. The approximation is obtained as a first derivative calculation on the sample space, using the derivative $\varphi(\theta)$; the approximating model is exponential and can be written:

$$\begin{aligned}\hat{f}(s; \theta) ds &= (2\pi)^{-p/2} e^{k/n} \exp \left[\ell(\theta) - \ell(\hat{\theta}^0) + \left\{ \varphi(\theta) - \varphi(\hat{\theta}^0) \right\} s \right] h(s) ds \\ &= (2\pi)^{-p/2} e^{k/n} \exp \left[-r^2 \{ \varphi(\theta), s \} / 2 \right] |\hat{J}_{\varphi\varphi}|^{-1/2} ds,\end{aligned}\tag{2.1}$$

where s is a nominal score variable to present the model, with $s^0 = 0$, $-r^2(\varphi, s)/2$ is the log likelihood ratio $\ell\{\theta(\varphi)\} - \ell\{\hat{\theta}(\hat{\varphi})\} + (\varphi - \hat{\varphi})s$ from the exponential model, and $\hat{\varphi} = \hat{\varphi}(s)$ is the maximum likelihood value from s in that exponential model. The approximate model is called the tangent exponential model and provides a full second order approximation to the original model in the moderate deviation range and a third order approximation to first derivative at the data point. It implements conditioning on an approximate ancillary which corresponds to the vectors V .

In §3 we will derive a p -value for testing the full p -dimensional parameter vector θ . For testing a sub-parameter ψ of length d , we can apply the same approximation, using the adjusted log-likelihood derived in Fraser (2003). Third order log-likelihood

for a scalar or vector parameter $\psi(\theta)$ can also be obtained from $\ell(\theta)$ and $\varphi(\theta)$ and has the form

$$\ell_{adj}(\psi) = \ell_P(\psi) + \frac{1}{2} \log |j_{[\lambda\lambda]}(\hat{\theta}_\psi)| \quad (2.2)$$

where λ is a complementing nuisance parameter, $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ is the constrained maximum likelihood value given $\psi(\theta) = \psi$, $\ell_P(\psi) = \ell(\hat{\theta}_\psi) = \ell(\psi, \hat{\lambda}_\psi)$ is the profile log-likelihood, and

$$|j_{[\lambda\lambda]}(\hat{\theta}_\psi)| = |j_{\lambda\lambda}(\hat{\theta}_\psi)| |\varphi_\lambda(\hat{\theta}_\psi) \hat{j}_{\varphi\varphi} \varphi'_\lambda(\hat{\theta}_\psi)|^{-1} \quad (2.3)$$

is the nuisance information determinant calculated relative to a symmetrized version of λ . Use of this log-likelihood to construct a conditional p -value is illustrated in the second example in §4.

3 A conditional p -value for inference about θ

We follow the approach developed in Fraser & Massam (1985), Skovgaard (1988) and Cheah *et al.* (1994), which considers the magnitude of the departure from a fixed value θ_0 , conditional on the direction of the departure. One approach for example might be to use the distribution of $|\hat{\theta} - \theta_0|$ conditional on $(\hat{\theta} - \theta_0)/|\hat{\theta} - \theta_0|$, but this is not parameterization invariant. In Cheah *et al.* (1994), where the model is a full exponential family in the canonical parametrization, the departure from θ_0 is measured by the score variable or sufficient statistic s , and an approximation is derived for the distribution of $|s - s_0|$, given $(s - s_0)/|s - s_0|$, where s_0 is the value of the score variable with maximum likelihood value θ_0 ; it is parametrization invariant.

Here we work directly with the tangent exponential model approximation (2.1). We denote by χ^2 the score based measure of departure from θ_0 using the tangent exponential model:

$$\chi^2 = \chi^2(\theta_0; s) = (s - s_0)' j_{\varphi\varphi}^{-1}(\theta_0) (s - s_0), \quad (3.1)$$

where s is the model variable with observed value $s^0 = 0$, and $s_0 = \ell_\varphi(\theta_0; y^0)$ is the expectation parameter value corresponding to θ_0 . Note that χ is proportional to $|s - s_0|$ for directional departure. The tangent exponential model (2.1), with $\theta = \theta_0$ can be expressed in terms of χ^2 as

$$\hat{f}(s; \theta_0) ds = (2\pi)^{-p/2} \exp \left\{ -\frac{\chi^2}{2} \right\} A(s) |j_{\varphi\varphi}(\theta_0)|^{-1/2} ds$$

where

$$A(s) = e^{k/n} \exp \left\{ \frac{\chi^2}{2} - \frac{r^2}{2} \right\} \left\{ \frac{|j_{\varphi\varphi}(\theta_0)|}{|\hat{j}_{\varphi\varphi}|} \right\}^{1/2}$$

and r^2 is the likelihood ratio statistic for testing $\theta = \theta_0$:

$$r^2 = r^2(\theta_0; s) = 2 \left[\ell(\hat{\theta}) - \ell(\theta_0) + \left\{ \varphi(\hat{\theta}) - \varphi(\theta_0) \right\} s \right]. \quad (3.2)$$

When r^2 is evaluated at the observed data point s^0 , it coincides with the likelihood ratio statistic from the original model $\ell(\theta; y^0)$. Both r^2 and χ^2 have a limiting χ_p^2 distribution under the model $f(y; \theta_0)$, and both are parametrization invariant.

We use the density approximation to compute the directional test of θ_0 conditioning on the score direction $e = (s^0 - s_0)/|s^0 - s_0| = -s_0/|s_0|$, where $s^0 = 0$ by the definition of the maximum likelihood estimate. The density of s is essentially a multivariate normal with an adjustment factor. Using this to compute the conditional distribution introduces a Jacobian effect s^{p-1} which combines with the normal kernel to give a chi density. The p -value based on this conditional test has the form

$$p(\theta_0) = \int_0^{x^0} g_d(\chi) \frac{a(\chi)}{c} d\chi$$

where c is the norming constant, $g_d(\chi)$ is the chi density used in (1.2), and

$$a(\chi) = A(\theta_0 + \chi e) \quad (3.3)$$

gives the adjustment factor along the line in the direction e .

It is shown in the Appendix that for functions $a(\chi)$ having an asymptotic expansion in χ , this p -value can be approximated to $O(n^{-1})$ by (1.1) where $\delta(\chi) = \chi^{-1} \{a(\chi)/a(0)\} - \chi^{-1}$ which in this case is

$$\delta(\chi) = \chi^{-1} \left[\exp \left\{ \frac{\chi^2}{2} - \frac{r^2}{2} \right\} \left\{ \frac{|j_{\varphi\varphi}(\theta_0)|}{|\hat{j}_{\varphi\varphi}|} \right\}^{1/2} - 1 \right] \quad (3.4)$$

Formula (3.4) has third order accuracy but its use in (1.1) yields just second order accuracy. The calculations in Andrews *et al.* (2005) indicate that third order accuracy for the directional p -value is not available from the usual asymptotic information $\{\ell(\theta), \varphi(\theta)\}$ and thus is not available without additional model information such as exponential or other specified model form (Cheah *et al.*, 1994).

With just second order accuracy available from (1.1), a reasonable option is to replace (3.4) by a further second order approximation:

$$\delta(\chi) = \left[\chi^2 - r^2 + \log \left\{ \frac{|j_{\varphi\varphi}(\theta_0)|}{|\hat{j}_{\varphi\varphi}|} \right\} \right] / 2\chi. \quad (3.5)$$

The likelihood and p -value expressions described here are easily seen to be invariant to the choice of parametrization, and invariant under re-expression of the response variable.

When (3.1), (3.2) and (3.5) are used in (1.1), they are evaluated at $s = s^0$.

The p -value has been developed in the context of inference for the full parameter vector θ . If we are interested in a sub-parameter vector $\psi(\theta)$ of dimension d , then the pair $\{\ell(\theta), \varphi(\theta)\}$ is replaced by the adjusted log-likelihood (2.2) and the corresponding likelihood gradient $\varphi(\psi)$. For scalar ψ this gradient is $\varphi(\psi) = \nu(\hat{\theta}) - \nu(\hat{\theta}_\psi)$ with ν calculated using a locally orthogonalized version of $\varphi(\theta)$. The p -value for assessing $\psi(\theta) = \psi$ depends only on the set $\Omega_\psi = \{\theta; \psi(\theta) = \psi\}$ and not on the structure of the parametrization $\psi(\theta)$ near $\psi(\theta) = \psi$. For vector ψ the gradient is determined component by component, using a version of ψ that has been orthogonalized at the observed data point.

4 Some examples

Example 1. Location normal: Consider $(y_1, \dots, y_p)'$ normally distributed with mean $(\theta_1, \dots, \theta_p)'$ and identity covariance matrix, and suppose the full parameter $\theta = (\theta_1, \dots, \theta_p)'$ is of interest. The log-likelihood for θ is

$$\ell(\theta; y) = -\frac{1}{2} \sum_1^p (y_i - \theta_i)^2,$$

both obviously and by the third order determination (Fraser, 2003). The related third order symmetrical canonical parametrization is $\varphi(\theta) = \theta$; and the related information is $j_{\varphi\varphi}(\theta_0) = I$. The likelihood ratio quantity (3.2) for assessing say θ_0 is

$$r^2 = \sum_1^p (s_i - s_{i0})^2 = \sum_1^p (y_i - \theta_{i0})^2$$

and the score based departure (3.1)

$$\chi^2 = (s - s_0)' I^{-1} (s - s_0) = \sum_1^p (y_i - \theta_{i0})^2.$$

It follows that $A(s) = 1$ and $\delta(\chi) = 0$, to second order. Thus from (1.1) we obtain

$$p(\theta_0) = G_p(\chi^0)$$

which is just the ordinary χ departure p -value but derived conditionally.

Example 2. Normal regression: Consider the normal regression model $y = X\beta + \sigma e$ where e is a sample from the standard normal. We assume that β has dimension r and thus $p = r + 1$, and we suppose that the parameter of interest is $\beta_{(1)} = (\beta_1, \dots, \beta_d)'$. The directional results are parametrization and sample space invariant, so here it suffices to take a canonical design matrix $X = (I \ O)'$. Let $s^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=r+1}^n y_i^2$ be the sum of squared residuals, $t = (\hat{\beta}_{(1)} - \beta_{(1)})/s$ be the Student statistic for the parameter of interest, and $T^2 = |t|^2 = \sum_1^d (y_i - \beta_i)^2 / s^2$ be the ratio of the usual sums of squares for inference about $\beta_{(1)}$. The third order likelihood $\beta_{(1)}$ is derived in Fraser (2003),

$$L(\beta_{(1)}) = (1 + T^2)^{-(n-f)/2} \quad (4.1)$$

where $f = r - d$ is the number of nuisance regression coefficients. This corresponds to the multivariate Student density for the usual multivariate quantity t for assessing $\beta_{(1)}$.

Location scale invariance shows (Fraser, 2003) that much of the analysis can without loss of generality be developed for a particular data point, with say, $\hat{\beta}_{(1)} = 0, s^2 = n$. A differential change in the data point with respect to the first d coordinates gives a corresponding translation of the likelihood in terms of the coordinates of $\beta_{(1)}$. From Cakmak, Fraser, & Reid (1994) we have that the asymptotic model determined at and to first derivative at a data point can to second order be approximated by a location model, and that the location parameter determination is unique. It follows that (4.1) records the relative density for that model and that the score variable is t and the score parameter is $\beta_{(1)}$. For this the directional assessment of $\beta_{(1)}$ is given by $|t|$ with a $\{dF/(n-p)\}^{1/2}$ distribution where F has the F distribution with d and $n-f$ degrees of freedom. The expression for $p(\beta_{(1)})$ from (1.1) then gives a second order approximation to that directional test.

5 Discussion

We have developed a simple directed departure measure (1.1) with (3.1) and (3.5) that provides a statistical measure of where the observed data point is relative to a vector parameter value. The measure combines the familiar likelihood ratio departure measure and the score departure measure and attains second order accuracy. The calculations are based on the log-likelihood and log-likelihood gradient quantities that are available in wide generality under moderate regularity conditions and a nominal assumption of asymptotic properties. Higher accuracy than the second order seems unavailable without additional information concerning the model, such as say explicit exponential model form. Some background on related likelihood theory may be found in §6 of Fraser (2003).

An alternative approximation for testing a vector parameter can be obtained by Bartlett correction of the likelihood ratio quantity (3.2). This is a scale correction, and the resulting approximate p -value is accurate to $O(n^{-2})$ (Barndorff-Nielsen & Hall, 1988). This higher accuracy comes from the restriction to likelihood ratio as the departure measure and is attained by sacrificing information specific to the direction of departure.

This is most obvious in the case of a scalar parameter of interest, where a Bartlett-corrected version is inherently two-sided, and when there is an underlying asymmetry can provide a very poor approximation to the p -value in the direction indicated by the data. As a simple scalar case where skewness is involved we take y to have the extreme value distribution with location parameter θ : $y = \theta + z$ and

$$f(z) = \exp\{-z - e^{-z}\}, \quad F(z) = \exp\{-e^{-z}\}.$$

The distribution is quite asymmetric; it has mode at 0 but the median is 0.366. The likelihood ratio quantity is

$$r^2 = 2 \left\{ \ell(\hat{\theta}) - \ell(\theta) \right\} = 2(z + e^{-z} - 1) \quad (5.1)$$

which has mean value 2γ , where $\gamma = 0.5772\dots$ is Euler's constant. Bartlett correction of the likelihood ratio leads to the confidence interval $(y^0 - 3.175, y^0 + 1.565)$; the exact confidence in the left tail is 0.85%, and in the right is 4.15%. The interval with exact confidence 2.5% in each tail is $(y^0 - 3.676, y^0 + 1.305)$. Both intervals are asymmetric around the point y^0 , but the equi-tailed interval seems more appropriate as a summary of the data. The approximation $\Phi(r^*)$ where $r^* = r + (1/r) \log(q/r)$, r is defined in (5.1), and $q = 1 - \exp(-z)$ provides very accurate inference in each tail, leading in this example to the third order 95% confidence interval $(y^0 - 3.699, y^0 + 1.309)$, with exact confidence 2.46% in the left tail and 2.44% in the right tail.

The anomaly is likely to be more pronounced in the case of vector parameters. As an illustration suppose y_1 and y_2 are independent observations from the extreme value distribution with location parameters θ_1 and θ_2 , respectively. The log-likelihood is

$$\ell(\theta; y) = -y_1 + \theta_1 - y_2 + \theta_2 - e^{(y_1 - \theta_1)} - e^{(y_2 - \theta_2)}$$

and

$$\varphi'(\theta) = \{\ell_{;y_1}(\theta; y^0), \ell_{;y_2}(\theta; y^0)\} = \{-1 + e^{(\theta_1 - y_1^0)}, -1 + e^{(\theta_2 - y_2^0)}\}$$

giving

$$\ell_\varphi(\theta; y^0) = \{1 - e^{-(y_1^0 - \theta_1)}, 1 - e^{-(y_2^0 - \theta_2)}\}'$$

and thus

$$\begin{aligned} \chi^2(\theta_0, s^0) &= \{e^{(\theta_{10} - y_1^0)} - 1\}^2 + \{e^{(\theta_{20} - y_2^0)} - 1\}^2 \\ r^2(\theta_0, y^0) &= 2\{-2 + y_1^0 - \theta_{10} + y_2^0 - \theta_{20} + e^{(y_1^0 - \theta_{10})} + e^{(y_2^0 - \theta_{20})}\}. \end{aligned}$$

Figure 1 shows the contours in the parameter space using the first order χ_2^2 approximation to the distribution of r^2 and the first order χ_2^2 approximation to the distribution of χ^2 , the Bartlett corrected version of r^2 , and the second order approximation using (1.1) and (3.5), all at the observed data point $(y_1^0, y_2^0) = (0, 0)$.

The Bartlett adjustment is a uniform rescaling of the likelihood ratio in all regions of the parameter space, whereas the adjustment to the score statistic is conditional on the direction. The second-order approximation here uses the standardized score function as the primary measure of departure, with an adjustment that is a function of the difference between that and the likelihood ratio. It should also be possible to start with the likelihood ratio and adjust it by some function related to the difference from the score. This is the approach taken in Skovgaard (2001), although Skovgaard approximates the needed sample space derivatives by expected values of various likelihood quantities; we compute the sample space derivatives using the tangent exponential model. Work on comparing the two methods in more complex examples is in progress.

6 Appendix

We want to evaluate

$$p(\theta_0) = \int_0^{\chi^0} g_d(\chi) \frac{a(\chi)}{c} d\chi$$

where c is the norming constant, $g_d(\chi)$ is the chi density used in (1.2), and

$$a(\chi)/a(0) = 1 + a_1\chi/n^{1/2} + a_2\chi^2/2n + O(n^{-3/2}).$$

Using $c = a(0)$ as a temporary norm to give say $\bar{p}(\theta_0)$, separating off the chi density and rewriting the discrepancy as

$$\delta(\chi) = \chi^{-1} \frac{a(\chi)}{a(0)} - \chi^{-1},$$

we obtain

$$\bar{p}(\theta_0) = G_d(\chi^0) + \int_0^{\chi^0} b_d g_{d+1}(\chi) \delta(\chi) d\chi,$$

where b_d is the ratio of χ norming constants recorded after (1.2). Since $\delta(\chi)$ is constant to $O(n^{-1})$ we obtain

$$\bar{p}(\theta_0) = G_d(\chi^0) + b_d G_{d+1}(\chi^0) \delta(\chi^0)$$

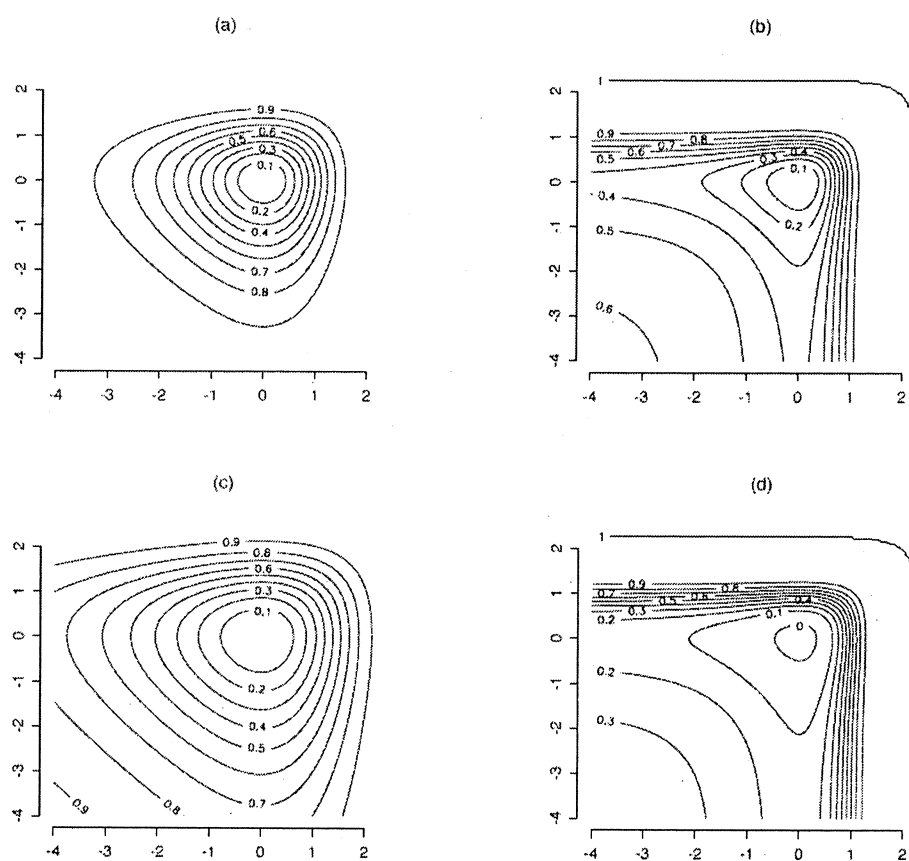


Figure 1. Parameter space contours for the bivariate extreme value distribution discussed in §5. These are based on: (a), the first order χ^2_2 approximation to the likelihood ratio statistic; (b), the first order χ^2_2 approximation to the score statistic; (c), the Bartlett-corrected likelihood ratio statistic; and (d), the second order approximation given by (1.1) using δ computed from (3.5). The observed data point is (0,0).

by integration by parts. Then adjusting for the temporary norm we obtain the second order p -value (1.1).

Acknowledgement

The Natural Sciences and Engineering Research Council of Canada has provided support for this research.

References

- [1] Andrews, D.F., Fraser, D.A.S., Li, R. & Wong, A. (2005). Distribution functions from local likelihood information. *J. Statist. Plan. Inf.* **134**, 130–193.
- [2] Barndorff-Nielsen, O.E. & Hall, P. (1988). On the level error after Bartlett adjustment of the likelihood ratio statistic. *Biometrika* **75**, 374–378.
- [3] Bartlett, M.S. (1955). Approximate confidence intervals III. A bias correction. *Biometrika* **42**, 201–4.
- [4] Cakmak, S., Fraser, D.A.S., & Reid, N. (1994). Multivariate asymptotic model: exponential and location approximations. *Utilitas Mathematica* **46**, 21–31.
- [5] Cheah, P.K., Fraser, D.A.S. & Reid, N. (1994). Multiparameter testing in exponential models: third order approximations from likelihood. *Biometrika* **81**, 271–8.
- [6] Fraser, D.A.S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327–339.
- [7] Fraser, D.A.S., & Massam, H. (1985) Conical tests: Observed levels of significance and confidence regions. *Statistische Hefte* **26**, 1–18.
- [8] Fraser, D.A.S. & Reid, N. (1993) Third order asymptotic models: Likelihood functions leading to accurate approximations for distribution functions. *Statist. Sinica* **3**, 67–82.
- [9] Fraser, D.A.S. & Reid, N. (1995). Ancillaries and third-order significance. *Utilitas Math.* **47**, 33–53.
- [10] Fraser, D.A.S. & Reid, N. (2001). Ancillary information for statistical inference. *Empirical Bayes and Likelihood Inference*. Ed. S.E. Ahmed and N. Reid, 185–207. New York: Springer.
- [11] Fraser, D.A.S., Reid, N. & Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–64.
- [12] Lugannani, R. & Rice, S. (1980). Saddlepoint approximation for the distribution function of the sum of independent variables. *Adv. Appl. Prob.* **12**, 47–90.
- [13] Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.* **31**, 1695–1731.
- [14] Skovgaard, I.M., (1988). Saddlepoint expansions for directional test probabilities. *J. R. Statist. Soc. B* **50** 269–80.
- [15] Skovgaard, I.M. (2001). Likelihood asymptotics. *Scand. J. Statist.* **28** 3–32.

Separated Families of Models: Sir David Cox Contributions and Recent Developments

M. I. Araujo

Depto. de Estatística, Universidade Federal do Amazonas, Brasil
E-mail: miaraujo@ufam.edu.br

B. de B. Pereira

Universidade Federal do Rio de Janeiro, Brasil
Email: basilio@ufrj.br

Robert Cleroux

University of Montreal, Canada
Email: robert.cleroux@umontreal.ca

Marcelo Fernandes

Graduate School of Economics, Getulio Vargas Foundation, Brasil
Email: m.fernandes@qmul.ac.uk

Aziz Lazraq

Ecole Nationale de l'Industrie Minerale, Morocco
Email: lazraq@enim.ac.ma

Abstract: This paper presents some recent classical and Bayesian procedures for choosing between separated multivariate linear regressions. It is shown that many of the recent results derive from suggestions in Sir David Cox (1961) fundamental paper and that, among the classical test procedures, only Cox test can be extendable to the multivariate case.

Key words: Bayes factors, Exponential compound model, Relative likelihood ratio, Separated families of hypotheses.

1 Introduction

Sir David Cox's contributions to the theory and application of statistics and stochastic processes indeed entail so much deepness and breadth that it is not surprising that many procedures and results bear his name: e.g., conditional inference counterexamples, sequential t-tests, large sample sequential tests, tests of separate families, Box-Cox transformation, supplementary variables in stochastic process, doubly stochastic processes, point processes, binary data, proportional hazard models, and graphical models.

In this paper, we focus on Cox's contributions to the hypothesis testing of separate families of models. We attempt to give a historical perspective, though we also present some of our recent results. Section 2 discusses the main contributions of Cox and their developments and extensions. Section 3 describes some Bayesian alternative solutions. Section 4 applies some of the testing procedures to systems of linear regression equations.

2 Separate family of hypothesis

The Neyman-Pearson theory of hypothesis testing applies only if the models belong to the same family of distributions. Special procedures are called for if the models belong to families that are separate (or nonnested) in the sense that an arbitrary member of one family cannot be obtained as a limit of members of the other. Let $y = (y_1, \dots, y_n)$ denote a random sample from some unknown distribution. The null hypothesis H_0 and alternative hypothesis H_1 respectively specify the parametric densities $f_0(y|\alpha_0)$ and $f_1(y|\alpha_1)$ for the random vector y , where α_0 and α_1 are unknown parameter vectors. Assume further that the density families are separate in the sense that arbitrary members of one family cannot be obtained as a limit of members of the other.

The asymptotic tests developed by Cox (1961, 1962) relies on a modification of the Neyman-Pearson likelihood ratio principle. The test statistic for H_0 against H_1 is

$$T_{01} = \Delta\ell_{01}(\hat{\alpha}_0, \hat{\alpha}_1) - n \left[\text{plim}_{n \rightarrow \infty} \frac{\Delta\ell_{01}(\hat{\alpha}_0, \hat{\alpha}_1)}{n} \right]_{\alpha_0 = \hat{\alpha}_0}, \quad (1)$$

where the plim is taken under H_0 , $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are respectively the maximum likelihood estimators of α_0 and α_1 , and $\Delta\ell_{01}(\alpha_0, \alpha_1)$ is the log-likelihood ratio $\ell_0(\alpha_0) - \ell_1(\alpha_1)$.

Cox showed that, asymptotically, T_{01} has a negative mean under the alternative hypothesis, whereas, under the null hypothesis, it is normally distributed with mean zero and variance

$$V_0(T_{01}) = V_0[\Delta\ell_{01}(\alpha_0, \alpha_{1|0})] - C_0' I_0^{-1} C_0, \quad (2)$$

where $\alpha_{1|0}$ is the probability limit of $\hat{\alpha}_1$ under H_0 , $C_0 \equiv n \frac{\partial}{\partial \alpha_0} \left[\text{plim}_{n \rightarrow \infty} \frac{1}{n} \Delta\ell_{01}(\hat{\alpha}_0, \hat{\alpha}_1) \right]$ and I_0 is the information matrix of α_0 . If one wishes to test H_1 against H_0 , analogous results hold for the statistic T_{10} . It then follows that $T_{01}^* = T_{01} V_0^{-1/2}(T_{01})$ and

$T_{10}^* = T_{10}V_1^{-1/2}(T_{10})$ are asymptotically standard normal under H_0 and H_1 , respectively. Note that the possible outcomes of the Cox test include inconclusive results such as the rejection, as well as the nonrejection, of both hypotheses, see Pereira (1998).

As an alternative, Cox (1961) proposed combining the two models in a general nesting specification of which they would both be special cases. Assuming that the density is proportional to the exponential mixture

$$[f_0(y|\alpha_0)]^\lambda [f_1(y|\alpha_1)]^{1-\lambda}, \quad (3)$$

it suffices to conduct inference about λ to test the relative merits of the two (or more) models. Despite its popularity in econometrics, the testing procedures based on the exponential mixture cannot be extended to simultaneous equations as shown by Pesaran (1982).

Another general approach suggested by Cox (1961) rests on Bayesian inference. The posterior odds for H_0 versus H_1 is

$$\frac{\pi_0}{\pi_1} \frac{q_0(y)}{q_1(y)} \equiv \frac{\pi_0}{\pi_1} B_{01}(y) \quad (4)$$

where π_j is the prior probability of H_j and

$$q_j(y) \equiv \int f_j(y|\alpha_j) \pi_j(\alpha_j) d\alpha_j \quad (5)$$

denotes the predictive distribution with prior probability $\pi_j(\alpha_j)$ for the parameters under H_j ($j = 0, 1$). The Bayes factor $B_{01}(y)$ represents the weight of evidence in the data favoring H_0 over H_1 . Cox also derived a general expression considering a loss function in the posterior odds, and described the large-sample approximation for the distribution of the Bayes factor. This approach has however two main limitations. First, the prior knowledge expressed by π_0 and $\pi_0(\alpha_0)$ must be coherent with that of π_1 and $\pi_1(\alpha_1)$. If the parameter spaces have different dimensions, for instance, there is no simple relation between the parameters. Second, if the prior information is weak and one applies an improper prior, the usual Bayes factor is not well defined. Section 3 discusses some alternatives that have been proposed so as to overcome these difficulties.

A series of unpublished PhD theses under Cox supervision at Imperial College put forth some further developments: e.g., O. A. Y. Jackson in 1968 and B. de B. Pereira in 1976 obtained further results on the modified likelihood ratio, A. C. Atkinson in 1970 developed the exponential compound model approach, and J. K. Lindsey in 1972 used a direct relative likelihood approach (see, for more details, Pereira 1977a). Further contributions of Cox in this field include Cox and Branwood (1959), Chambers and Cox (1967), Atkinson and Cox (1974), and Cox (1974). Since then a huge amount of research on separate families of hypothesis was developed. In the 1980s, the econometricians, using the exponential compound model, took a great interest in the subject. Bayesian statistician in the 1990's developed alternative Bayes factors. For reviews and references, see McAleer, Pesaran and Bera (1990), Gouriéroux and Monfort (1994), McAleer (1995), Pereira (1977b, 1981, 1998), and Pesaran and Weeks (2001).

3 Alternative Bayes factors

To overcome the difficulties with the usual Bayes factors, several modifications were proposed in the literature see Araujo and Pereira (2003). In what follows, we discuss the main alternative Bayes factors.

3.1 The posterior Bayes factor

Aitikin (1991) proposed the posterior Bayes factor that compares the posterior means of the likelihood function under H_0 and H_1 . More formally, the posterior density under H_j ($j = 0, 1$) is

$$\pi_j(\alpha_j|y) \equiv \frac{f_j(y|\alpha_j)\pi_j(\alpha_j)}{\int f_j(y|\alpha_j)\pi_j(\alpha_j) d\alpha_j}, \quad (6)$$

which yields

$$q_j^P(y) = \int f_j(y|\alpha_j)\pi_j(\alpha_j|y) d\alpha_j \quad (7)$$

as the posterior mean of the likelihood function under H_j ($j = 0, 1$). The posterior Bayes factor then corresponds to the ratio of the posterior means, namely $B_{01}^P(y) = q_0^P(y)/q_1^P(y)$.

3.2 The partial Bayes factor

O'Hagan (1991) derived a modification to Aitikin's (1991) posterior Bayes factor so as to avoid problems with improper priors. Consider the partition $y = (x, z)$ of the sample. From the subsample x , one obtains proper posterior densities $\pi_0(\alpha_0|x)$ and $\pi_1(\alpha_1|x)$ to use as the priors for the subsample z . The partial Bayes factor $B_{01}(z|x)$ then equals

$$B_{01}(z|x) \equiv \frac{\int f_0(z|x, \alpha_0)\pi_0(\alpha_0|x) d\alpha_0}{\int f_1(z|x, \alpha_1)\pi_1(\alpha_1|x) d\alpha_1} = \frac{q_0(y)/q_0(x)}{q_1(y)/q_1(x)} = \frac{B_{01}(y)}{B_{01}(x)}. \quad (8)$$

The idea is that improper priors presumably affect $B_{01}(y)$ and $B_{01}(x)$ in the same manner, and so their effects cancel out.

3.3 The intrinsic Bayes factor

Berger and Pericchi (1996) defined that the training sample x in the sample partition $y = (x, z)$ is minimal if the posteriors for α_0 and α_1 are proper and there is no subset of x that entails a proper posterior. There are usually many, say R , partitions featuring a minimal training sample. The idea of the intrinsic Bayes factor $B_{01}^I(y)$ is to take either the average or the median of the partial Bayes factors $\{B_{01}(z_r|x_r); r = 1, \dots, R\}$ obtained from the R minimal training samples. If R is too large, one may randomly select a sample from the collection of possible training samples. The arithmetic and

geometric intrinsic Bayes factors respectively are

$$B_{01}^{IA}(y) = \frac{1}{R} \sum_{r=1}^R B_{01}(z_r|x_r) \quad (9)$$

$$B_{01}^{IG}(y) = \left[\prod_{r=1}^R B_{01}(z_r|x_r) \right]^{1/R}, \quad (10)$$

whereas the median intrinsic Bayes factor is

$$B_{01}^{IM}(y) = \text{med}\{B_{01}(z_r|x_r); r = 1, \dots, R\}. \quad (11)$$

3.4 The fractional Bayes factor

O'Hagan (1995) proposed a modification of the Bayes factor that also relies on the use of a training sample. Let $b = n_x/n$ denote the training fraction, where n_x is the size of the training sample x . The fractional Bayes factor then is $B_{01}^{[b]}(y) = q_0^{[b]}(y)/q_1^{[b]}(y)$, with

$$q_j^{[b]}(y) = \frac{\int f_j(y|\alpha_j)\pi_j(\alpha_j) d\alpha_j}{\int [f_j(y|\alpha_j)]^b \pi_j(\alpha_j) d\alpha_j} \quad (12)$$

for $j = 0, 1$. To understand the fractional Bayes factor, observe that, for n_x large enough, the likelihood for the full sample is approximately equal to the power b of the likelihood for the training sample. It therefore is consistent provided that b shrinks to zero as n grows.

4 Multivariate linear regressions

4.1 The Cox likelihood ratio test

Consider two separate multivariate linear regression models $H_0 : Y = XB_0 + U_0$ and $H_1 : Y = ZB_1 + U_1$, where Y is a $n \times m$ matrix of regressands, X and Z are respectively $n \times p$ and $n \times q$ matrices of regressors, and B_0 and B_1 are respectively $p \times m$ and $q \times m$ matrices of parameters. The error terms U_0 and U_1 have rows that are independent and identically distributed as normal random vectors with mean zero and identity covariance matrices Σ_0 and Σ_1 , respectively. We also assume that X and Z are of full rank, with $n \geq m + p$ and $n \geq m + q$. It thus follows that: $U_0 \sim \mathcal{N}(0, I_n \otimes \Sigma_0)$ and $U_1 \sim \mathcal{N}(0, I_n \otimes \Sigma_1)$, whereas $Y \sim \mathcal{N}(XB_0, I_n \otimes \Sigma_0)$ under H_0 and $Y \sim \mathcal{N}(ZB_1, I_n \otimes \Sigma_1)$ under H_1 .

The matrices of regressors X and Z are fixed and nonnested in the sense that it is not possible to obtain the columns of X from the columns of Z , and vice versa. We further assume that the matrices $\Sigma_{X'X} \equiv \lim_{n \rightarrow \infty} \frac{1}{n} X'X$ and $\Sigma_{Z'Z} \equiv \lim_{n \rightarrow \infty} \frac{1}{n} Z'Z$ are nonsingular, and that $\Sigma_{X'Z} \equiv \lim_{n \rightarrow \infty} \frac{1}{n} X'Z$ is a nonzero matrix. The above assumptions ensure that the maximum likelihood estimators $\hat{B}_0 = (X'X)^{-1}X'Y$ and $\hat{B}_1 = (Z'Z)^{-1}Z'Y$ are consistent under H_0 and H_1 , respectively.

The Cox centered likelihood ratio test then is $\hat{T}_{01} = \frac{n}{2} (\log |\hat{\Sigma}_1| - \log |\hat{\Sigma}_{1|0}|)$, where $\hat{\Sigma}_1 = \frac{1}{n}(Y - Z\hat{B}_1)'(Y - Z\hat{B}_1)$, $\hat{\Sigma}_{1|0} = \hat{\Sigma}_0 + \frac{1}{n}\hat{B}_0'XM_ZX\hat{B}_0$, $\hat{\Sigma}_0 = \frac{1}{n}(Y - X\hat{B}_0)'(Y - X\hat{B}_0)$, and $M_Z = I - Z(Z'Z)^{-1}Z'$. For the variance of $\hat{T}_{01}$, see either Pesaran and Deaton (1978) or Davidson and MacKinnon (1983).

The multivariate case is only treated in Pesaran and Deaton (1978), Davidson and MacKinnon (1983), McAleer, Pesaran and Bera (1990), Timm and Al-Subaihi (2001). Unfortunately, the tests developed in the econometric literature cannot be extended for simultaneous equations as shown by Pesaran (1982). This is because tests based on the exponential mixture are not always consistent (Pereira, 1977). Only Davidson and MacKinnon (1983) mention the problem of identification risen by Pesaran (1982) that impedes the generalization of non nested hypotheses testing procedure other than Cox test to system of equations without imposing unrealistic assumptions.

4.2 Bayes factors

In this section, we extend the results of Aitikin (1991), O'Hagan (1995), and Berger and Pericchi (1996) to the context of multivariate linear regressions. The posterior odds for H_0 against H_1 is $(\pi_0/\pi_1)B_{01}$. Suppose that one uses improper priors for the parameters such that $\pi_0(\alpha_0)$ and $\pi_1(\alpha_1)$ are respectively proportional to constants K_0 and K_1 . The Bayes factor B_{01} then is proportional to K_0/K_1 and is not well defined. For the multivariate regression models, Jeffreys diffuse prior is given by

$$\pi_0(\alpha_0) = \pi_0(B_0)\pi_0(\Sigma_0) = K_0|\Sigma_0|^{-\frac{m+1}{2}}, \quad (13)$$

giving way to the following predictive distribution under the null hypothesis

$$q_0(Y) = \pi^{\frac{m(2n-2p-m+1)}{4}} K_0 |X'X|^{-m/2} |S_0|^{-\frac{n-p}{2}} \prod_{s=1}^m \Gamma\left(\frac{n-p-s+1}{2}\right), \quad (14)$$

where $S_0 \equiv (Y - X\hat{B}_0)'(Y - X\hat{B}_0)$. A similar expression holds for the alternative model $Y = ZB_1 + U_1$. The resulting Bayes factor then is

$$B_{01}(Y) = \pi^{m(p-q)/2} \frac{K_0}{K_1} \left(\frac{|Z'Z|}{|X'X|}\right)^{m/2} \frac{|S_1|^{(n-q)/2}}{|S_0|^{(n-p)/2}} \prod_{s=1}^m \frac{\Gamma\left(\frac{n-p-s-1}{2}\right)}{\Gamma\left(\frac{n-q-s-1}{2}\right)}, \quad (15)$$

where $S_1 \equiv (Y - Z\hat{B}_1)'(Y - Z\hat{B}_1)$. It is clear from (15) that the Bayes factor is not well defined for it depends on the unknown ratio K_0/K_1 .

From (14) and (15), it is now possible to derive the alternative Bayes factor. For instance, the posterior Bayes factor $B_{01}^P(Y)$ of Aitikin (1991) results from the ratio between

$$q_0^P(Y) = (2\sqrt{\pi})^{-mn} |S_0|^{-n/2} \prod_{s=1}^m \frac{\Gamma\left(\frac{2n-p-s+1}{2}\right)}{\Gamma\left(\frac{n-p-s+1}{2}\right)}$$

and

$$q_1^P(Y) = (2\sqrt{\pi})^{-mn} |S_1|^{-n/2} \prod_{s=1}^m \frac{\Gamma\left(\frac{2n-q-s+1}{2}\right)}{\Gamma\left(\frac{n-q-s+1}{2}\right)}.$$

It therefore ensues that

$$B_{01}^P(Y) = \left(\frac{|S_1|}{|S_0|} \right)^{n/2} \prod_{s=1}^m \frac{\Gamma\left(\frac{2n-p-s+1}{2}\right) \Gamma\left(\frac{n-q-s+1}{2}\right)}{\Gamma\left(\frac{2n-q-s+1}{2}\right) \Gamma\left(\frac{n-p-s+1}{2}\right)}. \quad (16)$$

The arithmetic intrinsic Bayes factor of Berger and Pericchi (1996) becomes

$$B_{01}^{IA}(Y) = \frac{1}{R} \sum_{r=1}^R \frac{B_{01}(Y)}{B_{01}(Y_{(r)})} = B_{01}(Y) \frac{1}{R} \sum_{r=1}^R B_{10}(Y_{(r)})$$

where $Y_{(r)}$ is a minimal training sample with design matrices $X_{(r)}$ and $Z_{(r)}$ under H_0 and H_1 , respectively. By definition, $Y_{(r)}$ is matrix such that both $X'_{(r)}X_{(r)}$ and $Z'_{(r)}Z_{(r)}$ are nonsingular. It has dimension $\bar{n} \times m$, where $\bar{n} = \lceil (m+1)/2 \rceil + \max(p, q)$ and $\lceil \cdot \rceil$ returns the smallest integer greater than its argument. From (15), it follows that

$$\begin{aligned} B_{01}^{IA}(Y) &= \left(\frac{|Z'Z|}{|X'X|} \right)^{m/2} \frac{|S_1|^{\frac{n-q}{2}}}{|S_0|^{\frac{n-p}{2}}} \prod_{s=1}^m \frac{\Gamma\left(\frac{n-p-s+1}{2}\right) \Gamma\left(\frac{\bar{n}-q-s+1}{2}\right)}{\Gamma\left(\frac{n-q-s+1}{2}\right) \Gamma\left(\frac{\bar{n}-p-s+1}{2}\right)} \\ &\quad \times \frac{1}{R} \sum_{r=1}^R \left(\frac{|X'_{(r)}X_{(r)}|}{|Z'_{(r)}Z_{(r)}|} \right)^{m/2} \frac{|S_{0(r)}|^{(\bar{n}-p)/2}}{|S_{1(r)}|^{(\bar{n}-q)/2}}, \end{aligned} \quad (17)$$

where $S_{j(r)}$ is analogous to S_j for the r -th minimal training set ($j = 0, 1$).

Finally, the fractional Bayes factor of O'Hagan(1995) results from the ratio between

$$q_0^{[b]}(Y) = \pi^{mn(1-b)/2} b^{mnb/2} |S_0|^{-n(1-b)/2} \prod_{s=1}^m \frac{\Gamma\left(\frac{n-p-s+1}{2}\right)}{\Gamma\left(\frac{nb-p-s+1}{2}\right)}$$

and

$$q_1^{[b]}(Y) = \pi^{mn(1-b)/2} b^{mnb/2} |S_1|^{-n(1-b)/2} \prod_{s=1}^m \frac{\Gamma\left(\frac{n-q-s+1}{2}\right)}{\Gamma\left(\frac{nb-q-s+1}{2}\right)}. \quad (18)$$

It then holds that

$$B_{01}^{[b]}(Y) = \left(\frac{|S_1|}{|S_0|} \right)^{n(1-b)/2} \prod_{s=1}^m \frac{\Gamma\left(\frac{n-p-s+1}{2}\right) \Gamma\left(\frac{nb-q-s+1}{2}\right)}{\Gamma\left(\frac{n-q-s+1}{2}\right) \Gamma\left(\frac{nb-p-s+1}{2}\right)}. \quad (19)$$

References

- [1] Aitkin, M. (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society B*, **53**, 111-142.
- [2] Araujo, M. I. and Pereira, B. B. (2003). Comparison among Bayes factors for separate models: some simulation results. Technical Report 2/2003, Centre for Multivariate Analysis, Department of Statistics, Penn State University.
- [3] Atkinson, A. C. and Cox, D. R. (1974). Planning experiments for discriminating between models (with discussion). *Journal of the Royal Statistical Society B*, **36**, 321-348.
- [4] Berger, J. O. and Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*. **91**, 109-122.

- [5] Cox, D. R. (1961). Tests of Separate Families of Hypotheses. *Proc. 4th Berkeley Symp.* **1**, 105-123.
- [6] Cox, D. T. (1962). Further Results on Tests of Separate Families of Hypotheses. *Journal of the Royal Statistical Society B.* **24**, 406-423.
- [7] Cox, D. R. (1974). Discussion of "Dempster, A.P. pg 353 - The direct use of likelihood for significance testing". In *Proceedings of Conference on Foundational Questions in Statistical Inference*. Memoirs n. 1, University of Aarhus.
- [8] Cox, D. R. and Brandwood, L. (1959). On a discriminatory problem connected with the works of Plato. *Journal of the Royal Statistical Society B.* **21**, 195-200.
- [9] Chambers, E. A. and Cox, D. R. (1967). Discriminating between alternative binary response models. *Biometrika.* **54**, 573-337.
- [10] Davidson, R. and MacKinnon, J. D. (1983). Testing the specification of multivariate models in the presence of alternative hypotheses. *Journal of Econometrics.* **23**, 301-313.
- [11] Gourieroux, C. and Monfort, A. (1994). Testing non-nested hypotheses. In *Handbook of Econometrics*. Vol. IV. Eds. R. Engle and D. L. McFadden. Elsevier, Ch. 44, 2585-2637.
- [12] McAller, M.; Pesaran, M.H. and Bera, A. K. (1990). Alternative approaches to testing non-nested models with autocorrelated disturbances. *Communications in Statistics - Theory and Methods.* **19**, 3619-3644.
- [13] McAller, M. (1995). The significance of testing empirical non-nested models. *Journal of Econometrics.* **65**, 149-171.
- [14] O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison (with discussion). *Journal of the Royal Statistical Society B.* **57**, 99-138.
- [15] Pereira, B. de B. (1977a). A note on the consistency and on finite sample comparisons of some tests of separate families of hypotheses. *Biometrika.* **64**, 109-113 (correction in volume 64, page 655).
- [16] Pereira, B. de B. (1977b) Discriminating among separate models: A bibliography. *International Statistical Reviews.* **45**, 163-172.
- [17] Pereira, B. de B. (1981). Discriminating among separate models: An additional bibliography. *International Statistical Information* 62(2), 3. In: Katti, S. K. - 1982 - On the Preliminary Test for the CEAS Model Versus the Thompson Model for Predicting Soybean Production. Technical Report 125, Department of Statistics, University of Missouri-Columbia.
- [18] Pereira, B. de B. (1998). Separate Families of Hypotheses. In *Encyclopedia of Biostatistics*. Ed. Armitage, P. and Colton, T. Vol. 5. John Wiley, pp. 4069-4074.
- [19] Pesaran, M. H. (1982). On the comprehensive method for testing non-nested regression models. *Journal of Econometrics.* **18**, 263-274.
- [20] Pesaran, M. H. and Deaton, A. S. (1978). Testing Non-nested Regression Models. *Econometrika.* **46**, 677-694.
- [21] Pesaran, M. H. and Weeks, M. (2001). Non-nested Hypothesis Testing: An Overview In *Companion to Theoretical Econometrics*. Ed. Badi H. Baltagi. Basil Blackwell. Oxford.
- [22] Timm, N. H. and Al-Subaihi A. A. (2001). Testing model specification in seemingly unrelated regression models. *Communications in Statistics - Theory and Methods.* **30**, 577-590.

LAD Regression and Nonparametric Methods for Detecting Outliers and Leverage Points

Susana Faria

*University of Minho, DMCT
P-4800-058 Guimarães, Portugal
E-mail: sfaria@mct.uminho.pt*

Giuseppe Melfi

*Université de Neuchâtel, Institut de Statistique
Espace de l'Europe 4
CH-2002 Neuchâtel, Switzerland
Email: Giuseppe.Melfi@unine.ch*

Abstract: The detection of influential observations for the standard least squares regression model is a question that has been extensively studied. LAD regression diagnostics offers alternative approaches whose main feature is the robustness. In this paper a new approach for nonparametric detection of influential observations in LAD regression models is presented and compared with other classical methods of diagnostics.

Key words: Least Absolute Deviations Regression, Robustness, Outliers, Leverage Points.

1 Introduction

The robustness of LAD to low-leverage outliers, and its susceptibility to high-leverage outliers has been extensively studied in literature [2,3,4]. In this paper we propose a method for nonparametric detection of such influential observations by the use of a technique derived from LAD regression. Robust methods based on the L_1 -norm have been proposed for example in [5,7]. The approach presented here considers suitable

perturbations of a given data set and allows a detection of high-leverage observations and outliers from a new viewpoint. These methods answer to natural requirements for robustness, and give a new tool for the analysis of data. Let $S \subset \mathbb{R}^{p+1}$ be a finite discrete set of points. In statistics, such a set may represent observations in $p + 1$ variables. Denote the elements of S as $(x_{i1}, \dots, x_{ip}, y_i)$, where the last variable is supposed to be explained from the preceding ones by a linear regression model:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i \quad \text{for } i = 1, \dots, n,$$

where p is the number of explanatory variables, ε_i are error terms, or deviations, and n is the number of observations. The LAD regression model is determined by minimizing the sum of the absolute deviations. In other words the vector of parameters $(\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}$ is determined by minimizing on $\beta_0, \beta_1, \dots, \beta_p$ the function

$$F(\beta_0, \beta_1, \dots, \beta_p) := \sum_{i=1}^n \left| y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right|. \quad (1)$$

When a linear LAD regression model is fitted, the hyperplane always passes through at least $p + 1$ points [1], although the solution may be non-unique.

For our purposes, we assume that for every dataset and for each subset of it which we deal with, the hyperplane which fits the linear LAD regression model is unique, as well as the observation with maximal absolute deviation. These assumptions are reasonable for datasets whose size is sufficiently large and/or the data are sufficiently precise in terms of significant digits. We suppose also that the dataset is such that every k points are not in the same $k - 1$ dimensional hyperplane, for $k = 3, 4, \dots, p + 2$. With these assumptions the linear LAD regression model is unique and it passes through exactly $p + 1$ points. Furthermore, if $n > p + 1$, there is always a point which does not belong to the regression hyperplane, so having a positive absolute deviation.

Consider the n datasets composed by all possible subsets of S of size $n - 1$. Under the above assumptions for each dataset we have a unique solution. For each case, we assign the score 1 to each point through which the fitted model passes and 0 to the other points. We define the final score of each point as the sum of scores over all models fitting the n datasets. This score is produced by the repeated use of the same points, each time considering a different subset of the original data set, so in a certain sense by bootstrapping the linear LAD regression model.

The point $(x_{k1}, x_{k2}, \dots, x_{kp}, y_k)$ will be also denoted by k and its score will be denoted by $L(k)$.

Similarly, we may define another complementary score function, denoted by $O(k)$, in the following way. Consider again the n datasets composed by all possible subsets of S of size $n - 1$. For each subset we consider the LAD regression line and we give the score 1 to the unique (according to the above assumptions) point which maximizes the absolute distance from the LAD regression line.

We define the score $O(k)$ as the sum (over all n possible subsets of S of size $n - 1$) of scores arising from the LAD regression lines.

In Section 2 we discuss some elementary properties of the LAD regression model. These properties will justify our algorithms for the detection of outliers and leverage points presented in Section 3. In Section 4 we discuss some examples, and compare the results with those obtained using other classical methods.

2 Preliminary Considerations

Under the above assumptions, the sum of L scores of all points is $n(p+1)$, and the sum of O scores is n , so, under the random variable viewpoint, $E(L(k)) = p+1$ and $E(O(k)) = 1$. Now suppose that we have a set of observations, all concentrated in a region and an isolated observation horizontally very far from the others but such that the line of the LAD regression model will pass through it (a typical leverage point). It is likely that the L score of this observation will be quite high. So a large L score is synonymous of leverage point. On the other hand, suppose we have a dataset in which all points are roughly in a hyperplane, and a further point far above them (an outlier). The LAD regression model will be very near to this hyperplane, and the score L of the outlier will be probably zero, but the score O will be probably $n-1$.

To justify these arguments we state the following theorems.

Theorem 1 *Let $(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)$ be n points in $\mathbb{R}^{p+1}$. Let $(x_{n+11}, \dots, x_{n+1p}, y_{n+1})$ be an additional point, such that $(x_{n+11}, \dots, x_{n+1p})$ belongs to the interior of the convex hull determined by the set $\{(x_{11}, \dots, x_{1p}), \dots, (x_{n1}, \dots, x_{np})\}$. If $x_{n+11}, \dots, x_{n+1p}$ are fixed and $|y_{n+1}|$ is sufficiently large, a hyperplane relative to a linear LAD regression model does not pass through $(x_{n+11}, \dots, x_{n+1p}, y_{n+1})$.*

Proof. Suppose $p = 1$ and that for $i = 1, \dots, n$, $c < y_i < d$. The convex hull hypothesis reduces to

$$a = \min_{i=1, \dots, n} \{x_i\} < x_{n+1} < \max_{i=1, \dots, n} \{x_i\} = b.$$

Let $y = \beta_0 + \beta_1 x$ be the line of the linear LAD regression model. Let ℓ_1 be the horizontal line $y = (c+d)/2$. The sum of the absolute deviations of $(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n+11}, \dots, x_{n+1p}, y_{n+1})$ does not exceed $(d-c)(n+1) + |y_{n+1}| + |d| + |c|$. So

$$F(\beta_0, \beta_1) < (d-c)(n+1) + |d| + |c| + |y_{n+1}|.$$

On the other hand, if a line $y = \beta_0^* + \beta_1^* x$ passes through (x_{n+1}, y_{n+1}) , if it is a linear LAD regression model, it will pass also through another point (x_i, y_i) for a suitable i , so for sufficiently large $|y_{n+1}|$, $|\beta_1^*| > |y_{n+1}|/(b-a)$. Hence there exists $\alpha > 1$ such that for sufficiently large $|y_{n+1}|$,

$$F(\beta_0^*, \beta_1^*) > \alpha |y_{n+1}|.$$

For sufficiently large $|y_{n+1}|$ we have

$$F(\beta_0^*, \beta_1^*) > \alpha |y_{n+1}| > (d-c)(n+1) + |d| + |c| + |y_{n+1}| > F(\beta_0, \beta_1),$$

and therefore, the line relative to the the LAD regression model cannot pass through (x_{n+1}, y_{n+1}) .

For $p > 1$ the proof is similar. S

Theorem 2 *Let $(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)$ be n points in $\mathbb{R}^{p+1}$. Let $(x_{n+11}, \dots, x_{n+1p}, y_{n+1})$ be an additional point. If y_{n+1} is fixed and $\sum_{i=1}^p |x_{n+1,i}|$ is sufficiently large, a linear LAD regression model will pass through $(x_{n+11}, \dots, x_{n+1p}, y_{n+1})$.*

Proof. The proof is an exercise, and the approach is similar to the proof of Theorem 1. S

Theorem 3 *Let $(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)$ be n points in $\mathbb{R}^{p+1}$, with $n > p + 1$. Let $L(k)$ and $O(k)$, for $k = 1, \dots, n$, defined as in Section 1. Then $L(k) + O(k) \leq n - 1$.*

Proof. This is a consequence of the fact that for each of the n subsets the scores are shared among distinct points, and, by the above assumptions, a point cannot collect a score for both L and O , since, for L it must have zero residual, and for O a strictly positive absolute residual. And each point appears exactly $n - 1$ times in the n subsets. S

3 The Algorithms

In this section we propose two algorithms based on the previous section. The aim of Algorithm 1 and Algorithm 2 is to detect leverage points and outliers respectively.

Algorithm 1.

- (1) Consider a data set S of size n . Let A and B be empty sets.
- (2) Let m be the size of S .
- (3) Consider all m subsets of S of size $m - 1$ and fit the LAD-regression model for each subset.
- (4) Compute $L(k)$ for each point of S and select $k_1 \in S$ which maximizes $L(k)$.
- (5) If $L(k_1) \geq \frac{8}{9}(m - 1)$ and $L(k_1) \geq \frac{3}{4}(n - 1)$ then move k_1 into B and move the eventual points of A into S ; otherwise move k_1 into A .
- (6) If the size of S does not exceed $\frac{9}{10}n$ then the process stops and B represents the leverage points; otherwise go to step 2.

In this algorithm, the elements of S are transferred in a set B of leverage points, or in a temporary set A where points that did not reached a sufficient score to be

classified as leverage points, are suitable to be reconsidered after that another point has been detected. This trick avoids the *masking effect*.

Discriminating values $\frac{8}{9}(m-1)$, $\frac{3}{4}(n-1)$ and $\frac{9}{10}n$ for the score function L , have been empirically determined, by testing on several data sets and several combinations of values. They have, however, a natural interpretation. When there is a unique leverage point, almost all m regression models detect it, so its L score is near to the maximum. When there are more leverage points, scores may be very different, and the *masking effect* can produce relatively small scores. Finally, we keep into account the size of S , to determine how many leverage points a data set may have. The process stops when the size of set S does not exceed $\frac{9}{10}n$, so with our method a data set cannot have more than $\frac{1}{10}n$ leverage points.

Algorithm 2.

- (1) Consider a data set S of size n . Let C and D be empty sets.
- (2) Initialize the last maximum score (LMS) by 0.
- (3) Let m be the size of S .
- (4) Consider all m subsets of S of size $m-1$ and fit the LAD-regression model for each subset.
- (5) Compute $O(k)$ for each point of S and select k_1 which maximizes $O(k)$.
- (6) If $O(k_1) = m-1$
 - a) Then if $O(k_1) = LMS-1$ or $LMS = 0$ then move k_1 into D , put $LMS = O(k_1)$ and move the eventual points of C into S ; otherwise the process stops and D represents the outliers.
 - b) Otherwise move k_1 into C .
- (7) If the size of S does not exceed $\frac{4}{5}n$ then the process stops and D represents the outliers; otherwise go to step 3.

In this algorithm, the set D contains the points classified as outliers and the set C contains the points that did not reached the score to be classified as outliers, and that are suitable to be reconsidered in further steps until the algorithm stops.

We can note the two proposed algorithms have a similar structure. However, the main difference is a feature of Algorithm 2: the outliers have decreasing scores O_1 , O_1-1 , O_1-2 , and so on. The algorithm stops when this sequence cannot be continued.

The process also stops when the size of set S does not exceed $\frac{4}{5}n$, so here a data set cannot have more than $\frac{1}{5}n$ outliers.

Discriminating values for the score function O , have been also empirically determined.

4 Some Examples

In this section we illustrate the proposed algorithms and compare them with other two methods using several real and simulated data sets.

One of the method is the P-R plot proposed by Hadi [5] to aid in classifying observations as leverage points, outliers or combinations of both. Some authors suggested that points with $h_{ii} > \frac{2(p+1)}{n}$, where h_{ii} is the i th diagonal element of matrix H , p is the number of predictors and n the number of observations, can be classified as leverage points and the points with $\frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}} > 2$, where r_i is the residual of the i th observation, h_{ii} is the i th diagonal element of matrix H and $\hat{\sigma}$ is an estimator of standard deviation of the errors, can be classified as outliers. In what follows the use of these suggested cut-off points to classify the observations will be intended as *classical methods*.

The first data set 'Telephone' relate the number of international telephone calls from Belgium (in tens of millions in minutes) to the variable year for 24 years and can be found in [7]. Cases 15-20 are unusually high and they are outliers. The second one 'Hawkins' consists of 75 observations in four dimensions, one response variable and three predictor variables, and can be found in [6]. It has been constructed for the study of special pathological phenomena in detection of outliers and leverage points and the cases 1-10 are outliers and leverage points. The data set 'Scottish' describes how the record times (in seconds) of 35 Scottish Hill races is related to two predictor variables, distance of race (in miles) and climb (in feet), and can be found in [5]. The data contain two clear outliers (observation 7 and 18). The last two data sets have been created by the authors. The data set 'twovariables' consists of 56 observations on one predictor variable and a response variable. The predictor was created as uniform $(0, 10)$ and the response variable to be consistent with the model $Y = X_1 + 4 + \epsilon$ with $\epsilon \sim N(0, 1)$. Three observations (51-53) have been conceived as leverage points and three others (54-56) as outliers. The other data set 'threevariables' is the three variable equivalent to the preceding one (two predictor variables).

Computation has been performed with a computer code in Splus, and the results are summarized in Table 1.

As we can see in Table 1, our proposed method performed very well in detecting all outliers in the data set 'Telephone'. The other methods failed to identify all of them because the observation 19 and 20 mask all the others.

In the case of 'Hawkins' data, our proposed method as well as the other two methods failed to identify the outliers. The outliers are all also swamped in the good cases 11-14. On the other hand, our method detected almost all leverage points.

In the data set 'Scottish', Table 1 shows that all three methods identified correctly the observation 7 and 18 as outliers. These observations mask the observation 33 detected by our proposed method. The observation 11 is suitably detected as leverage point by all three methods, but there are others observations identified as leverage points only in one or two methods.

Data	Method	Leverages	Outliers
Telephone	Classical Method	-	20
	Hadi's Method	-	19, 20
	Our Results	-	17-20
Hawkins	Classical Method	12-14	7, 11-14
	Hadi's Method	14	7, 11-14
	Our results	3-6, 9, 10, 13	11-14
Scottish	Classical Method	7, 11, 33, 35	7, 18
	Hadi's Method	7, 11	7, 18
	Our Results	11, 17, 35	7, 18, 33
Twovariables	Classical Method	51-53	54-56
	Hadi's Method	51-53	54-56
	Our results	52	54-56
Threevariables	Classical Method	18, 51-53	54-56
	Hadi's Method	51-53	54-56
	Our Results	51-53	9, 37, 54-56

Table 1: Detection of outliers and leverage points according to different methods.

For the simulated data sets, all methods performed very well in detecting all outliers. However, our proposed method failed to identify all leverage points.

5 Conclusion

The computation of the scores requires the determination of a certain number of LAD regression models, and this is computationally longer than usual methods. However it is important to note that the principle is very simple, and takes into account natural requirements for robustness in the detection of influential observations. Nowadays, the performances of common notebooks are largely sufficient to perform in a few seconds the computations for the above examples, so the new tools are suitable for applications in the statistical methodology. A code has been implemented in Splus and is available on demand.

References

- [1] Arthanari, T.S.; Dodge, Y., 1993. *Mathematical Programming in Statistics*, Classic edition, John Wiley and Sons, New York.
- [2] Dodge, Y. (Ed.) 1987, *Statistical data analysis based on the L_1 -norm and related methods*, Elsevier/North-Holland, New York; Amsterdam.
- [3] Dodge, Y. (Ed.) 1997, *L_1 -statistical procedures and related topics*, Institute of Mathematical Statistics, Hayward.

- [4] Dodge, Y. (Ed.) 2002, *Statistical Data Analysis Based on the L_1 -Norm and Related Methods*, Birkhäuser.
- [5] Hadi, A., 1992. A new measure of overall potential influence in linear regression, *Comp. Stat. and Data Analysis*, **14**, 1–27.
- [6] Hawkins, D.M., Bradu, D. and Kass, G.V., 1984. Location of several outliers in multiple regression data using elemental sets, *Technometrics*, **26**, 197–208.
- [7] Rousseeuw, P.J. and Leroy, A.M., 1987. *Robust Regression and Outlier Detection*, John Wiley and sons, New York.

On Partial LAD Regression

Athanassios Kondylis

*Université de Neuchâtel, Institute of Statistics
Espace de l'Europe 4
CH-2002 Neuchâtel, Switzerland
Email: Athanassios.Kondylis@unine.ch*

Abstract: The partial least absolute deviation regression seeks to model the median of the response instead of the mean which is the case for partial least squares (PLS) regression. An iteratively reweighted PLS algorithm based on LAD regression residuals is proposed which provides robust and fast results in high dimensional regression problems.

Key words: Median regression model, L_1 norm, LAD regression, partial least squares.

1 Introduction

In a regression problem there is usually a response vector $\mathbf{y}$ and a set of recorded covariates $\mathbf{x}_j$, with $j = 1, \dots, p$. For the recorded units $(\mathbf{x}_i, y_i)$, indexed by $i = 1, \dots, n$, we generally have

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \epsilon_i, \quad (1)$$

where $\boldsymbol{\beta}$ is a set of unknown parameters, f is the functional form which maps the predictors on the response, and ϵ_i is the random error.

The classical linear regression assumes function f to be linear on the parameters $\boldsymbol{\beta}$ and proceeds in estimating the conditional expectation of $\mathbf{y}$ given $\mathbf{X}$ as below

$$E(y_i | \mathbf{x}_i) = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \text{ and } \text{var}(y_i | \mathbf{x}_i) = \sigma^2, \quad (2)$$

where $\mathbf{X}$ is the design matrix which includes in its columns the covariates $\mathbf{x}_j$, $\hat{\boldsymbol{\beta}}$ is the estimated parameter vector, and σ^2 is the residual variance.

Partial least squares (PLS) regression builds linear regression models in terms of derived components instead of original predictors. It uses ordinary least squares to regress the response on the derived components [17]; let us denote the k^{th} component by t_k .

By using derived components instead of the original predictors the PLS regression reduces the dimension of the regression problem from a large number of variables to a small number of retained components $(t_1, \dots, t_k)$. Their number is often selected by means of cross-validation or bootstrap techniques [2]. PLS components are good predictors for the response since they are derived in directions $\mathbf{w}$ which solve

$$\max_{\mathbf{w}} \{ \text{cor}^2(\mathbf{X}\mathbf{w}, \mathbf{y}) \text{var}(\mathbf{X}\mathbf{w}) \}, \quad (3)$$

subject to $\mathbf{w}^T \mathbf{w} = 1$ and $\mathbf{X}\mathbf{w} \perp \mathbf{X}\mathbf{w}'$ for $\mathbf{w} \neq \mathbf{w}'$.

The notation cor and var above is used to denote sample correlation and variance for vectors, while $\perp$ indicates vector's orthogonality.

The use of ordinary least squares regression in the PLS regression algorithm allows to model the mean of the response conditional on the derived PLS components. The use of regression estimates other than the ordinary least squares has been already mentioned in the literature [4, 16]. With this paper we wish extend the use of LAD regression on PLS regression setting by modifying the partial LAD (PLAD) regression, as introduced in [4], by using an iteratively reweighted algorithm. The idea of iteratively reweighting the PLS regression algorithm is not new. It was used in [1] as well as in [6] and [11]. For additional robust PLS regression methods one can see [10], [16], as well as [12].

We main focus on two reasons for which we give an iteratively reweighted PLAD regression algorithm. Firstly, because the use of weights on both the Y and the X-space allows the proposed algorithm to be robust on both X and Y directions. Secondly, because we seek for smooth partial LAD estimates with faster computation. In what follows, we briefly review the univariate PLS and PLAD regression in Section 2. In Section 3, a smooth and robust algorithm for the partial LAD regression is given. Section 4 uses real and experimental data sets in order to illustrate and study the proposed method. Finally, in Section 5 conclusions are drawn.

2 PLS and PLAD regression

The bilinear regression model is described in [13]. It is the most natural representation of the bilinear decomposition of $\mathbf{X}$ and $\mathbf{y}$ that PLS regression carries out. This is based on the following operations:

1. The original data $(\mathbf{X}, \mathbf{y})$ are centered or standardised.
2. For $k \leq p$

- The covariance $\text{cov}(\mathbf{x}_{j,k-1}, \mathbf{y}_{k-1})$ is stored in the vector $\mathbf{w}_k$. The latter is pooled in the matrix $\mathbf{W}_{(p \times k)}$,
- Components $\mathbf{t}_k$ are extracted, whith $\mathbf{t}_k = \sum_{j=1}^p \mathbf{x}_{j,k-1} \mathbf{w}_k$.
- $\mathbf{X}_{k-1}$ and $\mathbf{y}_{k-1}$ are orthogonalised with respect to $\mathbf{t}_k$,
- Least squares residuals are taken as new data for $k \leftarrow k + 1$.

A small number of components ($k \ll p$) is usually sufficient to build the final regression model. This is commonly determined by means of cross validation [2]. For an excellent overview on PLS regression one can see [18].

The PLAD regression method is introduced by [4]. It uses the L_1 norm associated with LAD regression. It follows exactly the structure of the univariate PLS regression by extracting its components in directions that depend upon the response variable. These directions are determined by a Gnanadesikan-Ketterning covariance estimate [5] that replaces common variance with the median absolute deviation (mad) based on the L_1 norm.

Both the univariate PLS regression (PLS1) and the PLAD regression share the following properties

- (1) $\mathbf{y} = q_1 \mathbf{t}_1 + \dots + q_k \mathbf{t}_k + \boldsymbol{\epsilon}$,
- (2) $\mathbf{X} = \mathbf{p}_1 \mathbf{t}_1 + \dots + \mathbf{p}_k \mathbf{t}_k + \mathbf{f}$,
- (3) $\text{cov}(\mathbf{t}_i, \mathbf{t}_j) = 0$ for $i \neq j$,

with $\boldsymbol{\epsilon}$ and $\mathbf{f}$ corresponding to residual terms, while the vector $\mathbf{p}_k$ corresponds to the X-loadings vector. The first two properties arise from the bilinear factorization [13]

The third property is due to the fact that both regression methods deflate data using least squares residuals.

PLS1 and PLAD regression relate the predictors to the response according to the following model

$$\hat{\mathbf{y}}_k = \sum_{j=1}^p \hat{\beta}_j \mathbf{x}_j, \quad (4)$$

but for different predictions $\hat{\mathbf{y}}_k$ and estimates $\hat{\beta}_j$.

The two regression methods differ substantially on the regression coefficient vector $\hat{\mathbf{q}}_k$ which is estimated using either least squares or LAD regression. While the PLS regression method estimates regression coefficients by minimizing the smooth least squares objective function, the PLAD is based on the minimization of the following objective function

$$\sum_{i=1}^n |y_i - \mathbf{x}_i^T \beta|, \quad (5)$$

which corresponds to the LAD objective function. The latter is known to be robust against outliers in the Y -direction. Yet, the PLAD regression does not bound the influence arising from high leverages on the predictor's space. Therefore it is not always a robust alternative to PLS regression.

3 A robust PLAD regression algorithm

The LAD regression estimates are essentially M -estimates. Hence, they are easily restated as a reweighted least squares problem [9]. Naturally the PLAD algorithm is easily reformulated in an iteratively reweighted least squares routine as follows.

Let $\mathbf{h}_k^y = \{h_{i,k}^y\}$ denote the weight vector taken into account at each iteration k . It is defined at each k by the LAD scaled residuals of the (residual) response $\mathbf{y}_{k-1}$ on the derived components. It is computed according to

$$h_{i,k}^y = g\left(\frac{\epsilon_{i,k}}{s}\right), \quad (6)$$

with $g(\cdot)$ being either a zero-one or a smooth function which allows to downweight observations with large scaled residual $\epsilon_{i,k}/s$. The scale parameter s is commonly chosen to be either the Interquartile Range (IQR) or the Median Absolute Deviation (mad). We remind that $IQR = Q_3 - Q_1$ with Q_t denoting the t^{th} quartile, and

$$mad(\mathbf{z}) = \text{med}(|\mathbf{z} - \text{med}(\mathbf{z})|),$$

where $\text{med}(\mathbf{z})$ denotes the sample median of vector $\mathbf{z}$. The new algorithm is rather fast to compute since it runs the ordinary PLS regression algorithm by reweighting the observations with the weights in (6). For 0-1 weights it simply excludes the outlying observations from the analysis.

In order to make the algorithm robust on both Y and X directions, we should introduce weights on the predictors [3]. The weights on the predictors can be computed in two ways. In the first case, weights similar to expression (6) can be obtained after regressing each predictor to the response using the LAD regression. These are denoted as $h_i^{x_j}$ after we removed the subscript k for notational convenience. Final weights on X are then computed according to

$$h_i^x = \prod_{j=1}^p h_i^{x_j}. \quad (7)$$

In the second case weights on the predictors can be computed following

$$h_i^x = g(d_i), \quad (8)$$

for a properly chosen distance d_i . The choice of (7) makes the algorithm robust and the whole process is based on LAD. Yet, the choice of (8) makes the algorithm faster and guarantees components orthogonality. In both cases, the final weights are given as

$$h_i = h_i^x h_i^y,$$

and the smoothed PLAD is robust on both the X and the Y direction. For the $k = 1$ the weights are defined by using the distances d_i defined above on the X -space and the Y -space.

4 Experience with real and simulated data

Real and simulated data will be used in this section in order to test the proposed PLAD regression algorithm and to compare it to PLS regression. The real life data set result from climatological studies, while the simulated data follow the latent variable model setting. With these examples we wish to provide a short description of how the proposed algorithm can be used and interpreted.

Throughout our experimentation we use the PLAD algorithm with weights given by

$$h_{i,k}^y = \begin{cases} 1 & \text{if } |\epsilon_{i,k}^*| \leq c, \\ 0 & \text{otherwise,} \end{cases} \quad (9)$$

where $c = \text{med}(\epsilon_{i,k}^*) + 1.5 \cdot \text{IQR}(\epsilon_{i,k}^*)$, and the star exponent indicating scaled residuals. LAD residuals are scaled by their IQR . Finally, 0-1 weights (like above) are taken on the predictors according to distances from the medians given by

$$d_i = \frac{\|\mathbf{x}_i - \mathbf{m}\|}{\text{med}(\|\mathbf{x}_i - \mathbf{m}\|)}, \quad (10)$$

with $\mathbf{m}$ denoting the multivariate L_1 median [8], and $\|\cdot\|$ the vector norm. For the choice of ((10)) one can also see [16].

4.1 Climatological Data

The temperatures taken from thirty-five weather stations across Canada over a year (365 days) have been registered and are accessible through the FDA homepage in "<http://ego.psych.mcgill.ca/misc/fda/index.html>". The thirty-five weather stations represent Pacific, Continental, Atlantic, and Arctic climates. The temperatures for the latter are well distinguished from the rest of the stations. They represent the last lines in the bottom of Figure 1. For an illustration of the use of PLS regression on functional data one can see [15, 14].

We want to predict the temperature (y) for the 35 weather stations the 365th day of the year given the 35×364 predictors matrix with the temperatures for all stations. We run the PLS and the PLAD regressions and we compute the root mean squared error (RMSE) loss for each one using five-fold cross validation [7]. The RMSE loss

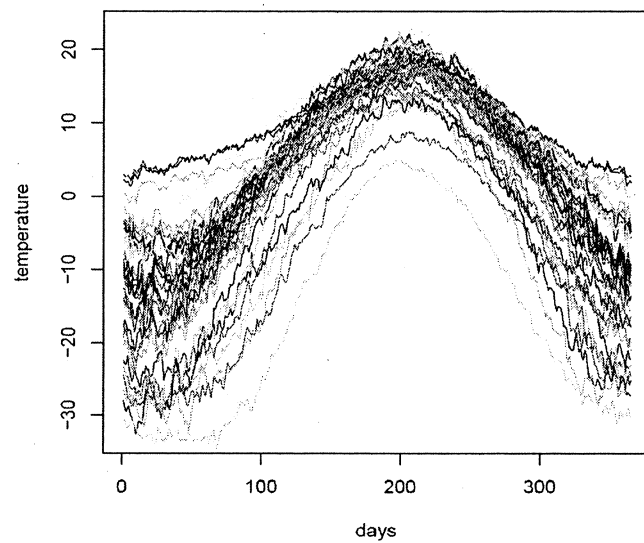


Figure 1. Temperature data. Temperatures over 365 days for 35 wether stations in Canada.

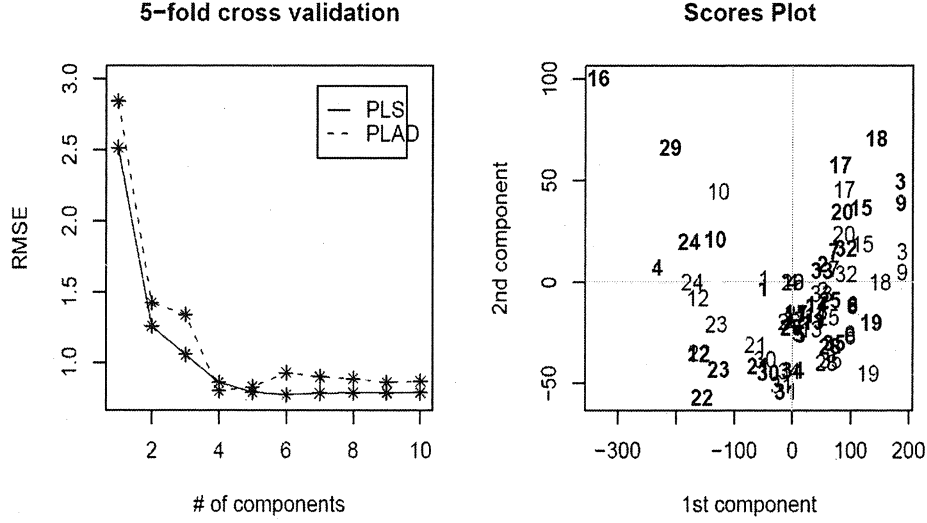


Figure 2. Temperature data. (a) *Left panel*: $\text{RMSE}(k)$ loss for PLS and PLAD regression for $k = 1, \dots, 10$. (b) *Right panel*: Score plot for the 1st and 2nd components. The bold numbers correspond to the PLS scores for the 35 wether stations.

is computed as

$$\text{RMSE} = \sqrt{\frac{1}{n_t} \sum_{i=1}^{n_t} (y_i - \hat{y}_i)^2},$$

for n_t denoting the size of the left out fold. The resulting lines are illustrated in the left panel of Figure 2.

PLS regression undoubtedly retains five components in the final regression model. The PLAD regression minimizes its loss on four components. This is probably due to the fact that extreme values arising mainly from arctic weather stations exert influence on PLS regression estimates which recover only on the fourth component. This is verified by inspecting the right panel of Figure 2. The scores plot on the first two components reveals wether stations which lie far away from the rest of the stations. These correspond to the arctic stations. The PLAD regression excludes these observations from the construction of the scores by setting their weights to 0.

4.2 Experimental data

We simulate here from a parameter vector β which is either a functionnal or a random normal variate. In the latter case it is simulated according to $\beta \sim \mathcal{N}(\mu, \sigma)$, with μ commonly set to zero. In the former case the regression coefficient is given by

$$\beta = 1 + 0.5 \sin(z) + \cos(0.5z), \text{ for } z \in [-4; 4], \quad (11)$$

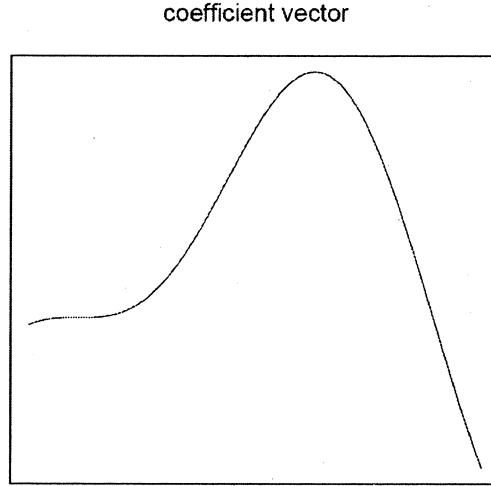


Figure 3. Experimental data. Implied regression coefficient in a functional form.

that is illustrated in Figure 3.

Given the parameter vector β we simulate from

$$\mathbf{y} = \sum_{j=1}^p \beta_j \mathbf{x}_j + \epsilon, \quad (12)$$

where the last term is taken either as standard normal or as laplace random variate. The latter incorporates longer variation on the response. In order to compute the response we need the data matrix $\mathbf{X}$. This is calculated according to the latent structure $\mathbf{X} = \mathbf{T}\mathbf{W}^T$, where $\mathbf{T} \sim \mathcal{N}_k(0, 1)$, and $\mathbf{W}$ is an arbitrary orthonormal $p \times k$ matrix.

We are interested to study the expected loss of the regression coefficient vector. This is computed, for both PLS and PLAD regression, in terms of its mean squared error ($\text{MSE}(\hat{\beta}_k)$) given by

$$\text{MSE}(\hat{\beta}_k) = M^{-1} \sum_m (\hat{\beta}_k^m - \beta)^T (\hat{\beta}_k^m - \beta). \quad (13)$$

The subscript k denotes the number of components on which the estimate was obtained, and $m = 1, \dots, M$ indicates the simulation run. We simulate 500 data sets and we measure the relative efficiency between the PLS and the PLAD regression vectors. This is given by the ratio

$$\text{RE.MSE}(\hat{\beta}_k) = \frac{\text{MSE}(\hat{\beta}_k^{\text{pls}})}{\text{MSE}(\hat{\beta}_k^{\text{plad}})}. \quad (14)$$

The simulation results are given in Table 1. The results are tabulated for all $k = 1, \dots, 5$. The out of sample prediction error was computed throughout the simulations and indicated a relative minimum on two components for both PLS and PLAD regression methods.

Table 1: Relative efficiencies ($\text{RE.MSE}(\hat{\beta}_k)$) for PLS and PLAD regression coefficients. The measures are averaged over the 500 simulated data sets.

$\beta = 1 + 0.5 \sin(z) + \cos(0.5z)$					
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Normal error	1.0004	1.0191	0.8750	1.5440	1.6011
Laplace error	1.3025	3.5340	5.8996	12.0468	11.9388
$\beta \sim (\mu, \sigma)$					
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Normal error	0.9999	0.9952	0.9703	1.1914	1.1997
Laplace error	1.0051	1.7020	2.9122	11.0363	13.5275

For this simulation setting, the results given in Table 1 demonstrate the efficiency of the PLAD method when the error term is normal. For Laplace error distribution the PLAD method yields better results than PLS regression in the sense that the $\text{RE.MSE}(\hat{\beta}_k) > 1$. PLS regression performance is even worse when we simulate from a functional parameter β (see the upper part of Table 1).

5 Conclusions

Within the context of PLS regression we used derived components in order to model the median of the response instead of the mean. This was achieved by using an iteratively reweighted least squares algorithm with weights based on LAD scaled residuals. Real world experiments from climatology demonstrated that in certain cases the PLAD regression corrects for outlying observations which influence the regression results. Finally, experience arising from simulations revealed that PLAD regression coefficients are efficient when the error term is normal, yet, they limit their loss compared to PLS regression estimates when longer variation in the error term is incorporated.

References

- [1] Cummins, D. and Andrews, C. (1995). Iteratively reweighted partial least squares: A performance analysis by monte carlo simulation. *Journal of Chemometrics*, 9:489–507.
- [2] Denham, M. (2000). Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *Journal*

- of *Chemometrics*, 14:351–361.
- [3] Dodge, Y. (1997). LAD regression for detection of outliers in response and explanatory variables. *Journal of Multivariate Analysis*, 61:144–158.
 - [4] Dodge, Y., Kondylis, A., and Whittaker, J. (2004). Extending PLS1 to PLAD regression and the use of the L1 norm in soft modelling. In *COMPSTAT 2004*, pages 935–942. Physica/Verlag - Springer.
 - [5] Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28:81–124.
 - [6] Griep, M. I., Wakeling, I. N., Vankeerberghen, P., and Massart, D. L. (1995). Comparison of semirobust and robust partial least squares procedures. *Chemometrics and Intelligent Laboratory Systems*, 29:37–50.
 - [7] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *Elements of Statistical Learning*. Springer, New York.
 - [8] Hössjer, O. and Croux, C. (1995). Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. *J. Nonparam. Statist.*, 4:293–308.
 - [9] Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.
 - [10] Hubert, M. and Branden, K. (2003). Robust methods for Partial Least Squares Regression. *Journal of Chemometrics*, 17:537–549.
 - [11] Kondylis, A. (2003). *Partial Reweighted Least Squares*. Master Thesis, University of Neuchâtel-Switzerland.
 - [12] Kondylis, A. and Hadi, A. S. (2006). Derived Components Regression using the BACON algorithm. *Computational Statistics & Data Analysis*, to appear.
 - [13] Martens, H. and Naes, T. (1989). *Multivariate Calibration*. Wiley, UK.
 - [14] Preda, C. and Saporta, G. (2005a). Clusterwise PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 49:99–108.
 - [15] Preda, C. and Saporta, G. (2005b). PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 48:149–158.
 - [16] Serneels, S., Croux, C., Filzmoser, P., and Van Espen, P. (2005). Partial Robust M-regression. *Chemometrics and Intelligent Laboratory Systems*, 79:55–64.
 - [17] Tenenhaus, M. (1998). *La régression PLS. Théorie et pratique*. Technip, Paris.
 - [18] Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130.

Bayesian Estimation of the GARCH(1,1) Model with Normal Innovations

David Ardia

*Séminaire d'Économétrie
Département d'Économie Quantitative
Université de Fribourg, Boulevard de Pérolles 90
CH-1700 Fribourg, Switzerland
Email: david.ardia@unifr.ch*

Abstract : In this article, we propose the Bayesian estimation of the parsimonious but effective GARCH(1,1) model with Normal innovations. We sample the parameters joint posterior distribution using the approach suggested by Nakatsuma [9]. As a first step, we fit the model to foreign exchange log-returns time series and compare the Maximum Likelihood and the Bayesian estimates. Next, we illustrate some appealing aspects of the Bayesian approach through interesting probabilistic statements made on the parameters.

Key words: GARCH model, Bayesian estimation, Markov Chain Monte Carlo.

1 Introduction

Volatility plays a central role in financial risk management and lies at the heart of any model for pricing derivatives securities. Research on changing volatility using time series models has been active since the pioneer paper by Engle [4]. From there, ARCH and GARCH type models grew rapidly into a rich family of empirical models for volatility forecasting during the 80's. These stochastic processes, unconditionally non-Gaussian and possibly non-stationary, have been extensively applied to financial data in the econometric literature. Reasons for that lying in their ability to reproduce heteroscedasticity, volatility clustering and unconditional heavy-tailed distribution, both salient features for reproducing financial returns.

Whilst apparently simple by nature, we may encounter many difficulties when dealing with GARCH models; the model's parameters must be positive, and sometimes the model is required to be covariance stationary, which can complicate the optimiza-

tion procedure. In addition, the finite sample evidence on the performance of GARCH Maximum Likelihood estimates and test statistics is still fairly limited. Reliable inference from the LM, Wald and LR test statistics generally require moderately large sample sizes of at least two hundred or more observations. However, most of these deficiencies break down when taking a Bayesian point of view. Any constraints on the parameters can easily be integrated in the Markov Chain Monte Carlo (MCMC) procedure, exact inference in small samples is possible and non-nested models can be tested using Bayes factors. Furthermore, distributions of complex functions of the parameters can be obtained by simulation at low cost in contrast to the bootstrap approach.

In this article, we propose the Bayesian estimation of the parsimonious but effective GARCH(1,1) model with Normal innovations. We sample the parameters joint posterior distribution using the approach suggested by Nakatsuma [9]. As a first step, we fit the model to foreign exchange log-returns time series and compare the Maximum Likelihood and the Bayesian estimates. Next, we illustrate some appealing aspects of the Bayesian approach through interesting probabilistic statements made on the parameters.

The plan of this article is as follows: we set up the model in Section 2. The MCMC scheme is detailed in Section 3. The empirical results are presented in Section 4 while we give some illustrative applications of the Bayesian approach in Section 5. We conclude in Section 6.

2 The model and the priors

A GARCH(1,1) model with Normal innovations may be written as:

$$\begin{aligned} y_t &= \varepsilon_t h_t^{1/2} \quad \text{for } t = 1, \dots, T \\ \varepsilon_t &\stackrel{iid}{\sim} N(0, 1) \\ h_t &:= \alpha_0 + \alpha_1 y_{t-1}^2 + \beta h_{t-1} \end{aligned} \tag{1}$$

where $\alpha_0 > 0$, $\alpha_1 \geq 0$ and $\beta \geq 0$; $N(0, 1)$ is the standard Normal distribution. In this setting, the conditional variance h_t is a linear function of the squared past observation and the past variance. Positivity restrictions on the parameters ensure a positive conditional variance. In order to write the likelihood function, we define the following vectors: $\mathbf{y} := (y_1 \cdots y_T)'$, $\boldsymbol{\alpha} := (\alpha_0 \ \alpha_1)'$ and we regroup the parameters into $\Theta := (\boldsymbol{\alpha}, \beta)$. In addition, we define the $(T \times T)$ diagonal matrix $\Sigma := \Sigma(\Theta) = \text{diag}(\{h_t(\Theta)\}_{t=1}^T)$ where:

$$h_t(\Theta) = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta h_{t-1}(\Theta) \quad .$$

From there, the likelihood function of Θ can be written as:

$$l(\Theta|\mathbf{y}) \propto (\det \Sigma)^{-1/2} \exp \left[-\frac{1}{2} \mathbf{y}' \Sigma^{-1} \mathbf{y} \right] \tag{2}$$

where, for convenience, we use the first observation as an initial condition and the initial variance is fixed to α_0 . This likelihood refers to the conditional likelihood of

the GARCH process given in (1). We propose the following proper priors on the parameters α and β of the preceding model:

$$\begin{aligned} p(\alpha) &\propto N_2(\alpha|\mu_\alpha, \Sigma_\alpha) \mathbb{I}_{[\alpha>0]} \\ p(\beta) &\propto N(\beta|\mu_\beta, \Sigma_\beta) \mathbb{I}_{[\beta>0]} \end{aligned}$$

where $\mu_\bullet$ and $\Sigma_\bullet$ are the hyperparameters, $\mathbb{I}_{[\bullet]}$ is the indicator function which equals unity if the constraint holds, and zero otherwise, and N_d is the d -dimensional Normal distribution ($d > 1$). In addition, we assume prior independence between α and β which implies that $p(\Theta) = p(\alpha)p(\beta)$. Then, we construct the joint posterior distribution via Bayes' rule:

$$p(\Theta|\mathbf{y}) \propto l(\Theta|\mathbf{y})p(\Theta) \quad (3)$$

3 Simulating the joint posterior

The recursive nature of the variance equation in (1) does not allow for conjugacy between the likelihood function and the prior distribution in (3). Therefore, we rely on the Metropolis-Hastings (M-H) algorithm to draw samples from the joint posterior distribution. The algorithm in this section is a special case of the algorithm described by Nakatsuma [9]. We draw an initial value $\Theta^{[0]} := (\alpha^{[0]}, \beta^{[0]})$ from the joint prior distribution and we generate iteratively J passes for Θ . A single pass is decomposed as follows:

$$\begin{aligned} \alpha^{[j]} &\sim p(\alpha|\beta^{[j-1]}, \mathbf{y}) \\ \beta^{[j]} &\sim p(\beta|\alpha^{[j]}, \mathbf{y}) \end{aligned}$$

Since no full conditional distribution is known analytically, we sample α and β from two proposal distributions. These distributions are obtained by noting that the GARCH(1,1) model can be written as an ARMA(1,1) model for $\{y_t^2\}$. In effect, by defining $w_t := y_t^2 - h_t$, we can transform the expression of the conditional variance as follows:

$$\begin{aligned} h_t &= \alpha_0 + \alpha_1 y_{t-1}^2 + \beta h_{t-1} \\ \Leftrightarrow y_t^2 &= \alpha_0 + (\alpha_1 + \beta) y_{t-1}^2 - \beta w_{t-1} + w_t \end{aligned} \quad (4)$$

where w_t can be written as:

$$w_t := y_t^2 - h_t = \left(\frac{y_t^2}{h_t} - 1 \right) h_t = (\chi_1^2 - 1) h_t$$

By construction, $\{w_t\}$ is a Martingale Difference process with variance $2h_t^2$ since the conditional expectation of w_t with respect to $\mathcal{F}_{t-1}$ (the natural filtration up to time $t-1$) is zero and the χ_1^2 variable has a unit mean and a variance equal to 2. However, as noted by Nakatsuma [9], it is difficult to generate Θ directly from equation (4). Hence, we approximate w_t by a variable z_t which is Normally distributed with a mean of zero and a variance of $2h_t^2$. This leads to the following *auxiliary* model:

$$y_t^2 = \alpha_0 + (\alpha_1 + \beta) y_{t-1}^2 - \beta z_{t-1} + z_t$$

By noting that z_t and h_t are both functions of Θ , respectively given by:

$$\begin{aligned} z_t(\Theta) &= y_t^2 - \alpha_0 - (\alpha_1 + \beta)y_{t-1}^2 + \beta z_{t-1}(\Theta) \\ h_t(\Theta) &= \alpha_0 + \alpha_1 y_{t-1}^2 + \beta h_{t-1}(\Theta) \end{aligned} \quad (5)$$

and by defining the $(T \times T)$ diagonal matrix $\Lambda := \Lambda(\Theta) = \text{diag}(\{2h_t^2(\Theta)\}_{t=1}^T)$ and the vector $\mathbf{z} := (z_1 \cdots z_T)'$ we can approximate the likelihood function of Θ from the auxiliary model as follows:

$$l(\Theta|\mathbf{y}) \propto (\det \Lambda)^{-1/2} \exp \left[-\frac{1}{2} \mathbf{z}' \Lambda^{-1} \mathbf{z} \right] \quad (6)$$

As will be shown hereafter, the construction of the proposal distribution for α and β is based on this likelihood function.

3.1 Generating α : ARCH coefficients

Recursive transformations initially proposed by Chib and Greenberg [3] allow to express the function $z_t(\Theta)$ in (5) as a linear function of α . Let us define $v_t := y_t^2$ for notational convenience. The recursive transformations are defined as follows:

$$\begin{aligned} l_t^* &:= 1 + \beta l_{t-1}^* \\ v_t^* &:= v_{t-1} + \beta v_{t-1}^* \end{aligned}$$

where the initial values l_0^* and v_0^* are set to zero. Let us regroup the terms within vectors: $\mathbf{v} := (v_1 \cdots v_T)'$, $\mathbf{c}_t := (l_t^* \ v_t^*)'$ and construct the $(T \times 2)$ matrix C where the t th row is $\mathbf{c}_t$. It turns out that $\mathbf{z} = \mathbf{v} - C\alpha$ and that we can express the likelihood function of α as follows:

$$l(\alpha|\beta, \mathbf{y}) \propto (\det \Lambda)^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{v} - C\alpha)' \Lambda^{-1} (\mathbf{v} - C\alpha) \right]$$

The proposal distribution to sample α is obtained by combining this likelihood function and the prior distribution by the usual Bayes update:

$$\begin{aligned} q_\alpha(\tilde{\alpha}, \alpha) &\propto N_2(\alpha | \hat{\mu}_\alpha, \hat{\Sigma}_\alpha) \mathbb{I}_{[\alpha > 0]} \\ \hat{\Sigma}_\alpha^{-1} &:= C' \tilde{\Lambda}^{-1} C + \Sigma_\alpha^{-1} \\ \hat{\mu}_\alpha &:= \hat{\Sigma}_\alpha (C' \tilde{\Lambda}^{-1} \mathbf{v} + \Sigma_\alpha^{-1} \mu_\alpha) \end{aligned}$$

where the $(T \times T)$ diagonal matrix $\tilde{\Lambda} := \text{diag}(\{2\tilde{h}_t^2\}_{t=1}^T)$ with:

$$\tilde{h}_t := \tilde{\alpha}_0 + \tilde{\alpha}_1 y_{t-1}^2 + \beta \tilde{h}_{t-1}$$

The value $\tilde{\alpha}$ is the previous draw of α in the M-H sampler. A candidate α^* is sampled from this proposal distribution and accepted with probability:

$$\min \left\{ \frac{p(\alpha^*, \beta | \mathbf{y})}{p(\tilde{\alpha}, \beta | \mathbf{y})} \frac{q_\alpha(\alpha^*, \tilde{\alpha})}{q_\alpha(\tilde{\alpha}, \alpha^*)}, 1 \right\}$$

3.2 Generating β : GARCH coefficient

The function $z_t(\Theta)$ in (5) could be expressed, in the previous section, as a linear function of α but can not be expressed as a linear function of β . To overcome the

problem, we linearize $z_t(\beta)$ by the first order Taylor expansion at point $\tilde{\beta}$, that is:

$$z_t(\beta) \simeq z_t(\tilde{\beta}) + \left. \frac{dz_t}{d\beta} \right|_{\beta=\tilde{\beta}} (\beta - \tilde{\beta})$$

where $\tilde{\beta}$ is the previous draw of β in the M-H sampler. Furthermore, let us define the following:

$$\begin{aligned} r_t &:= z_t(\tilde{\beta}) + \tilde{\beta} \nabla_t \\ \nabla_t &:= - \left. \frac{dz_t}{d\beta} \right|_{\beta=\tilde{\beta}} \end{aligned}$$

where the terms ∇_t can be computed by the following recursion:

$$\nabla_t := y_{t-1}^2 - z_{t-1}(\tilde{\beta}) + \tilde{\beta} \nabla_{t-1}$$

with the initial value $\nabla_0 := 0$. Then, we regroup these terms into vectors: $\mathbf{r} := (r_1 \cdots r_T)'$, $\mathbf{\nabla} := (\nabla_1 \cdots \nabla_T)'$ and we approximate the exponential in (6) by:

$$\exp \left[-\frac{1}{2} (\mathbf{r} - \beta \mathbf{\nabla})' \Lambda^{-1} (\mathbf{r} - \beta \mathbf{\nabla}) \right] .$$

The proposal distribution to sample β is obtained by combining the approximated likelihood and the prior distribution by Bayes' update:

$$\begin{aligned} q_\beta(\tilde{\beta}, \beta) &\propto N(\beta | \hat{\mu}_\beta, \hat{\Sigma}_\beta) \mathbb{I}_{[\beta > 0]} \\ \hat{\Sigma}_\beta^{-1} &:= \mathbf{\nabla}' \tilde{\Lambda}^{-1} \mathbf{\nabla} + \Sigma_\beta^{-1} \\ \hat{\mu}_\beta &:= \hat{\Sigma}_\beta (\mathbf{\nabla}' \tilde{\Lambda}^{-1} \mathbf{r} + \Sigma_\beta^{-1} \mu_\beta) \end{aligned}$$

where the $(T \times T)$ diagonal matrix $\tilde{\Lambda} := \text{diag}(\{2\tilde{h}_t^2\}_{t=1}^T)$ with:

$$\tilde{h}_t := \alpha_0 + \alpha_1 y_{t-1}^2 + \tilde{\beta} \tilde{h}_{t-1} .$$

A candidate β^* is sampled from this proposal distribution and accepted with probability:

$$\min \left\{ \frac{p(\beta^*, \boldsymbol{\alpha} | \mathbf{y}) q_\beta(\beta^*, \tilde{\beta})}{p(\tilde{\beta}, \boldsymbol{\alpha} | \mathbf{y}) q_\beta(\tilde{\beta}, \beta^*)}, 1 \right\} .$$

We end this section with some comments regarding the implementation of the MCMC scheme. The program is written in the R language with some subroutines written in C in order to speed up the simulation procedure. The validity of the algorithm as well as the correctness of the computer code are verified by a variant of the method proposed by Geweke [7]. We sample Θ from a proper joint prior and generate some passes of the M-H algorithm; at each pass, we simulate the dependent variable $\mathbf{y}$ from the full conditional $p(\mathbf{y} | \Theta)$ (given by the conditional likelihood). This way, we draw a sample from the joint distribution $p(\mathbf{y}, \Theta)$. If the algorithm were correct, the resulting replications of Θ should reproduce the prior. The Kolmogorov-Smirnov comparison test does not reject this hypothesis at the 1% level.

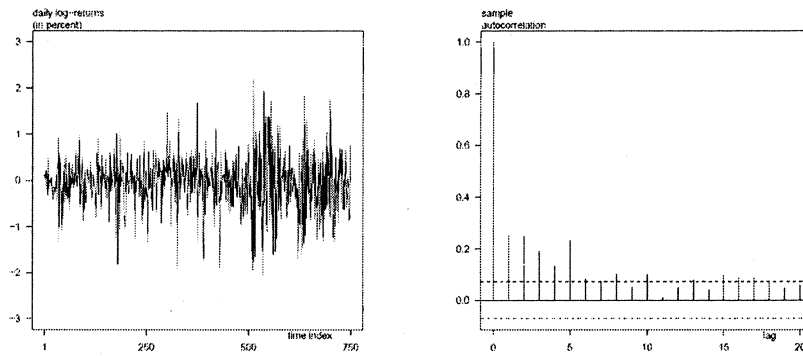


Figure 1. Daily log-returns (left) and sample autocorrelogram (right)

4 Empirical analysis

We apply our Bayesian estimation to daily observations of the Deutschmark vs British Pound foreign exchange log-returns. The sample period is from January 3, 1985, to December 31, 1991, for a total of 1974 observations. The nominal returns are expressed in percent. This data set has been promoted as an informal benchmark for GARCH time series software validation and is available from the Journal of Business and Economic Statistics (JBES). From this time series, the first 750 observations, which is somewhat less than three financial years, are used to illustrate the Bayesian approach. The number of data is large enough to perform classical Maximum Likelihood (ML) estimation and apply asymptotic justifications. Hence, we have an interesting point of view from which to compare classical and Bayesian approaches.

The observation window excerpt from our data set is plotted on the left-hand side of Figure 1. We test for autocorrelation in the times series by testing the joint nullity of autoregression coefficients for $\{y_t\}$. We estimate the regression with autoregression coefficients up to lag 20 and compute the covariance matrix using the White estimate. The p -value of the Wald test is 0.377 which does not support the presence of autocorrelation. However, from Figure 1 we clearly observe clusters of high and low variability in the time series. This phenomenon is well known in financial data and is referred to as volatility clustering. This effect is emphasized on the right-hand side of Figure 1 where the sample autocorrelogram of squared observations is displayed. In this case, the first autocorrelations are large and significant, indicating GARCH effects; the Wald test strongly rejects absence of autocorrelation in the squares.

As an additional data analysis, we test for unit root using the Phillips and Perron [11] test. The test strongly rejects the $I(1)$ hypothesis. From this preliminary analysis, we conclude that the time series is not integrated and does not exhibit autocorrelation. However, we strongly suspect the presence of GARCH effects in the data.

	Θ_{MLE}	$\bar{\Theta}$	$\Theta_{0.5}$	$\Theta_{0.025}$	$\Theta_{0.975}$	min	max	IF
α_0	0.039 [0.014,0.064]	0.048 (0.448)	0.047	0.022	0.080	0.011	0.119	9.79
α_1	0.198 [0.102,0.294]	0.226 (1.284)	0.223	0.128	0.337	0.083	0.499	5.85
β	0.686 [0.538,0.833]	0.636 (5.021)	0.636	0.476	0.795	0.338	0.849	40.79

Table 1: Estimation results for the GARCH(1,1) model. Θ_{MLE} : Maximum Likelihood estimate. $\bar{\Theta}$: posterior mean. Θ_ϕ : estimated posterior quantile at probability ϕ . IF: inefficiency factor (ratio of the variance of mean relative to a *iid* sequence). $[\bullet]$: Maximum Likelihood 95% confidence interval.

4.1 Model estimation

We fit the parsimonious GARCH(1,1) model to the data for this observation window. As a prior distribution for the Bayesian estimation we choose a truncated tri-dimensional Normal distribution with a zero mean vector and a diagonal covariance matrix. The variances are set to 10'000 so we do not introduce tight prior information into our estimation. We run two chains for 10'000 passes each. We emphasize the fact that only positivity constraints are implemented in the M-H algorithm; no stationarity conditions are imposed in the simulation process.

In Figure 2, the running mean is plotted over iterations. For all parameters, we notice a convergence of the two chains toward a constant value after something like 5'000 iterations. The diagnostic test by Gelman and Rubin [5] does not reject convergence of the chain after 5'000 passes (values ranging from 1.04 to 1.05 for the 97.5th percentile of the potential scale reduction factor). The one lag autocorrelations in the chains range from 0.75 for α_1 to 0.95 for β which is reasonable. The sampling algorithm allows to reach very high acceptance rates ranging from 89% for α to 95% for β . From the overall MCMC output, we discard the first 5'000 draws as a *burn in* period and merge the two chains to get a final sample's length of 10'000. In addition, we estimate the model by the usual Maximum Likelihood technique for comparison purposes.

The posterior statistics as well as the ML results are presented in Table 1. First, we note that even if the number of observations is large, the ML estimates and the Bayesian posterior means are different; the ML estimation is lower for α and higher for β . We also notice a difference between the 95% confidence intervals. Whereas the confidence band is symmetric in the ML case due to the asymptotic Normality assumption, this is not true for the posterior confidence intervals. The reason can be explained through Figure 3 where the posterior distributions for the parameters are displayed.

Indeed, we clearly notice the asymmetric shape of the histograms for α_0 and α_1 . The skewness values are 0.46 and 0.39, both significantly different from zero. Therefore the ML confidence band has a tendency to underestimate the right boundary of the

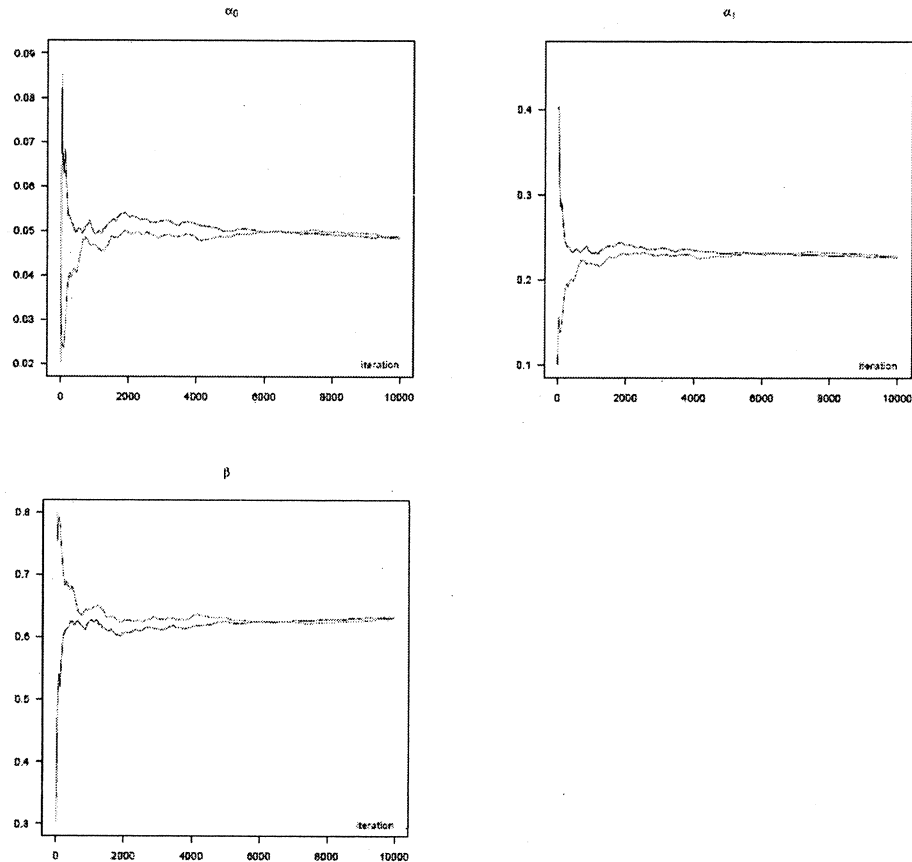


Figure 2. Running means of the chains over iterations (up to 10'000). The sampler generates α and β from the candidate distribution derived in Section 3. The acceptance rate ranges from 89% for α to 95% for β . The autocorrelations range from 0.75 for α_1 to 0.95 for β . The convergence diagnostic indicates convergence of the chains from iteration 5'000. The 97.5th percentile of the potential shrink factor ranges from 1.04 to 1.05.

95% confidence interval for these parameters. In the case of β , the skewness is -0.09 , also significant; in this case, the Maximum Likelihood overestimate the left boundary of the 95% confidence band. Furthermore, as shown in the bottom right-hand side of Figure 3, the joint distribution for the parameters α_0 and β is slightly different from the ellipsoid we should obtain under the multivariate Normal assumption. Therefore, these results warn us against the abusive use of asymptotic justifications; in the present case, even 750 observations do not suffice to assume the asymptotic Normal distribution for the parameters.

Finally, the last column of Table 1 gives the inefficiency factors for the different parameters. Their values are computed as the ratio of the squared numerical standard error of the MCMC simulations and the variance estimate divided by the number of iterations (i.e. the variance of the sample mean from a hypothetical *iid* sampler). The numerical standard errors are estimated by the method of Andrews [1], using a Parzen kernel and AR(1) pre-whitening as presented in Andrews and Monahan [2]. This ensures easy, optimal, and automatic bandwidth selection. Using 10'000 simulations out of the posterior distribution seems appropriate if we require that the Monte Carlo error in estimating the mean is smaller than one percentage of the variation of the error due to the data. The larger inefficiency factor reported for β is reflected in a larger autocorrelation in the simulated values.

4.2 Sensitivity analysis

The Bayesian approach is often criticized by the fact that the prior distribution for the parameters can have a significant impact to the posterior distribution, and as a consequence, bias the results. It is therefore important to determine the extend of this impact through a sensitivity analysis. As noted by Geweke [8, Section 2], it is possible to approximate the Bayes factor between two models differing only by their prior densities using the posterior simulation output from just one of the models. This approach provides an attractive way of performing sensitivity analysis since it does not require the estimation of the alternative model.

We test the sensitivity by considering four alternative prior distributions; either by modifying the mean or/and increasing the variances relative to our initial prior. The Bayes factors are then estimated as explained in the previous paragraph and ranked with the Jeffrey's scale of evidence. In all cases, we conclude to a weak evidence for our initial specification relative to the alternative prior, indicating that our initial prior is vague enough. The results are not shown to save space but can be obtained from the author upon request.

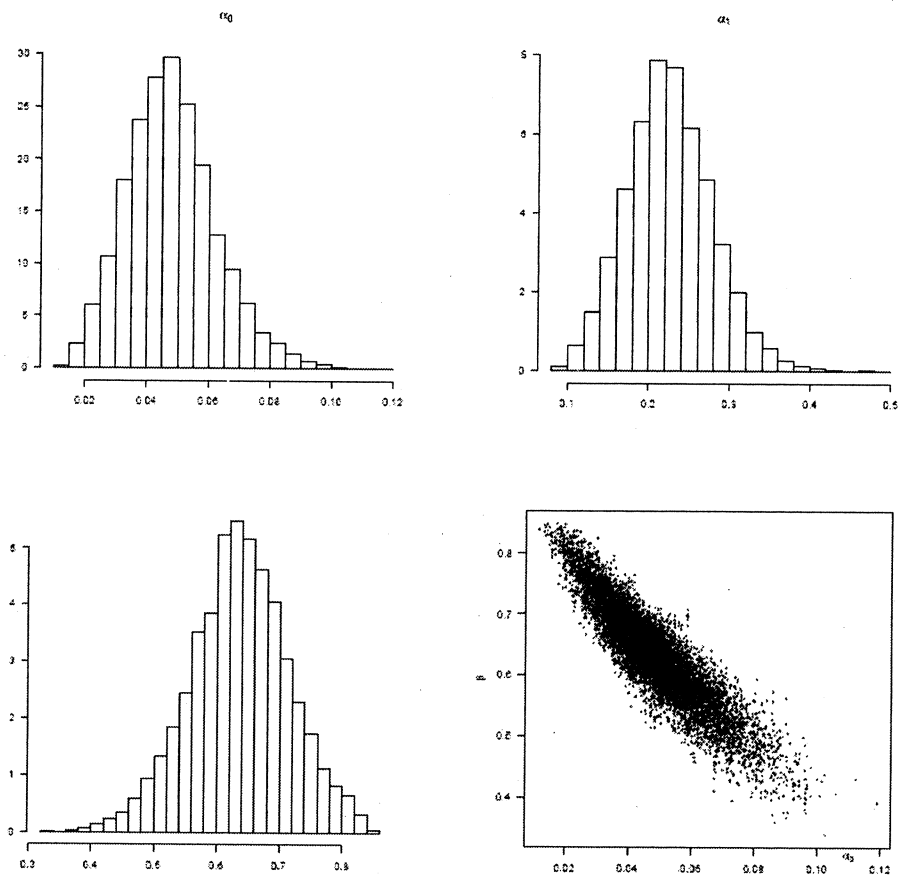


Figure 3. Posterior distributions for the GARCH(1,1) parameters based on 10'000 draws. In the lower-right graphic, we present a scatter plot of posterior (α_0, β) .

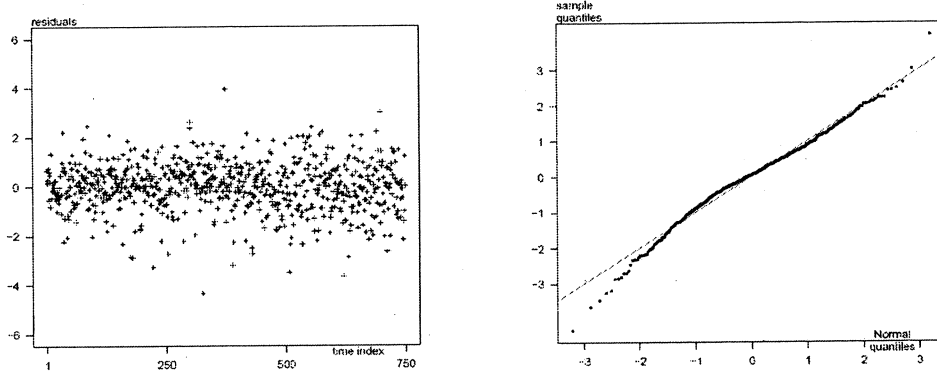


Figure 4. Residuals (left) and quantile-quantile plot (right)

1.3 Model diagnostic

We check for model misspecification by testing the standardized residuals.¹ They are defined by:

$$\hat{\varepsilon}_t := y_t \hat{h}_t^{-1/2}$$

for $t = 1, \dots, 750$ where $\hat{h}_t$ is the conditional variance computed with $\Theta_{0.5}$ (the median of the posterior sample). If the statistical assumptions in (1) are satisfied, these residuals should be independent and Normally distributed.

On the left-hand side of Figure 4 we display the residuals over time. No autocorrelation or heteroscedasticity are visually apparent. We test for autocorrelation using the Ljung-Box test up to lag 20. The test does not reject the null at the 5% level (p -value = 0.652). This is also true for the squared residuals (p -value = 0.961). Hence, the GARCH(1,1) process has been able to *filter* the heteroscedastic nature of the data. We form a quantile-quantile plot of the residuals against the Normal distribution on the right-hand side of Figure 4. The distribution is almost Normal at its center whereas the tails are slightly fatter, especially the left one. The Kolmogorov-Smirnov Normality test rejects the null hypothesis at the 5% level (p -value = 0.008). The tails of the innovations' distribution are not fat enough to fully capture the distributional nature of the data. This point is recurrent with financial data and heavy tails distributions for the innovations are sometimes useful to overcome this problem.

¹ An alternative would be to test the predictive performance of the model over an out-of-the-sample window (i.e. specification test). This approach is however not pursued in this article since we consider a single model and we focus on the estimation instead of the forecasting performance. A misspecification test is simpler and certainly sufficient in this context.

5 Illustrative applications

In this section, we illustrate some interesting probabilistic statements made possible under the Bayesian framework. The joint posterior sample is used to simulate complex functions of the parameters.

5.1 Persistence

As pointed out in Section 3, a GARCH(1,1) process for $\{y_t\}$ is equivalent to an ARMA(1,1) process for $\{y_t^2\}$ with an autoregressive coefficient $(\alpha_1 + \beta)$, and a moving average coefficient $-\beta$. The autocorrelation function (ACF) comes from the standard formulae for the ARMA(1,1) model. It is recursively given by $\rho_i := (\alpha_1 + \beta) \cdot \rho_{i-1}$ for $(i > 2)$ with the first order autocorrelation given by:

$$\rho_1 := \alpha_1 \frac{1 - \beta^2 - \alpha_1 \beta}{1 - \beta^2 - 2\alpha_1 \beta}$$

The term $(\alpha_1 + \beta)$ is the degree of persistence in the autocorrelation of the squares. It controls the intensity of the clustering in the variance process. With a value close to one, past shocks and past variances will have a longer impact on the future conditional variance. An autoregressive coefficient $(\alpha_1 + \beta) = 1$ corresponds to a unit root process for squared observations.

To make inference on persistence and ACF, we simply use the posterior sample $\Theta^{[j]}$ and generate $(\alpha^{[j]} + \beta^{[j]})$ and $\rho_i^{[j]}$ for $j = 1, \dots, 10'000$ and $i = 1, \dots, 20$. The graphic of the posterior distribution of the persistence $(\alpha_1 + \beta)$ is plotted on the left-hand side of Figure 5. The histogram is slightly left-skewed with a median value of 0.865 and a maximum value of 0.992. In this case, the integration for the variance process is not supported by the data. On the right-hand side of Figure 5 we display the posterior ACF with its 95% and 99% confidence bands together with the sample autocorrelations. Although a single observation (at lag 11) lies outside the confidence bands, the autocorrelation structure of the estimated GARCH(1,1) model is in line with the data.

5.2 Stationarity

In the case of the GARCH(1,1) process, Nelson [10] gave the conditions for covariance stationarity (CSC) and strict stationarity (SSC). These conditions are given by:

$$CSC := \alpha_1 + \beta - 1 < 0$$

$$SSC := \mathbb{E}[\ln(\alpha_1 \varepsilon_t^2 + \beta)] < 0$$

where the error term ε_t is Normally distributed. As pointed out in Section 4, the covariance stationary condition has not been imposed in the M-H algorithm. The joint posterior sample can be used to estimate the posterior distribution of these functions:

$$CSC^{[j]} := \alpha_1^{[j]} + \beta^{[j]} - 1$$

$$SSC^{[j]} := \frac{1}{K} \sum_{k=1}^K \ln(\alpha_1^{[j]} (\eta^{[k]})^2 + \beta^{[j]})$$

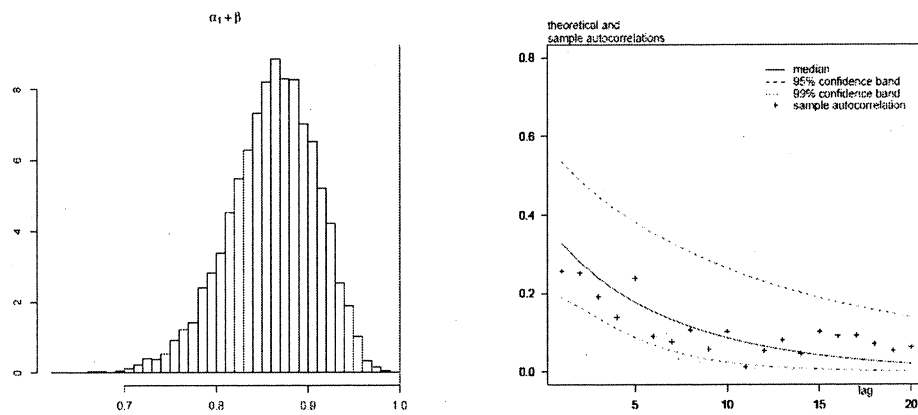


Figure 5. Posterior distribution for the persistence (left) and posterior autocorrelogram (right). The solid line is the posterior median, the dashed lines the 95% confidence bands and the dotted lines the 99% confidence bands. The cross symbols are values of the sample autocorrelation of the squared log-returns up to lag 20.

for $j = 1, \dots, 10'000$, where $\eta^{[k]}$ is a draw from a standard Normal distribution and K is set large enough (in our application we choose $K = 1'000$). In Figure 6 we present the posterior distributions for *CSC* and *SSC*. None of these values exceed zero in our simulation study. The estimated model is therefore covariance stationary and strictly stationary.

Other probabilistic statements on interesting functions can be obtained using the joint posterior sample. For example, the posterior median is 0.341 for the marginal variance and 4.54 for the marginal kurtosis. They approximately correspond to the sample estimations of 0.323 and 4.63.

6 Conclusion

This paper has proposed the estimation of the Bayesian GARCH(1,1) model with Normal innovations. The MCMC scheme has been derived in order to simulate the joint posterior distribution for the model's parameters. The GARCH(1,1) model has been applied to foreign exchange log-returns time series and comparison with the traditional Maximum Likelihood has been performed. It has been shown that even if the sample size is fairly large (in our case 750 observations), point estimates differ slightly between the two approaches. In addition, the posterior distribution for some parameters is skewed which warn us against the abusive use of the Normal approximation. A sensitivity analysis has been performed in order to robustify the estimation results. Finally, we have illustrated some appealing aspects of the Bayesian approach through interesting probabilistic statements made on the parameters.

As a final comment, we note that some financial models might use GARCH parameters as input quantities. This is the case for instance with the Black-Scholes formula in options pricing, which is a function of the marginal variance of the underlying financial asset. Under GARCH(1,1) dynamics, this marginal variance is $\alpha_0/(1 - \alpha_1 - \beta)$ if the process is covariance stationary, and this value can be simulated to obtain a full distribution. Then, the posterior distribution of the marginal variance could be used to simulate the option price's distribution. Furthermore, subjective constraints on the parameters could be integrated in the MCMC procedure; for instance, an option trader could set prior lower or/and upper boundaries for the marginal variance and then run the estimation process to estimate the GARCH parameters.

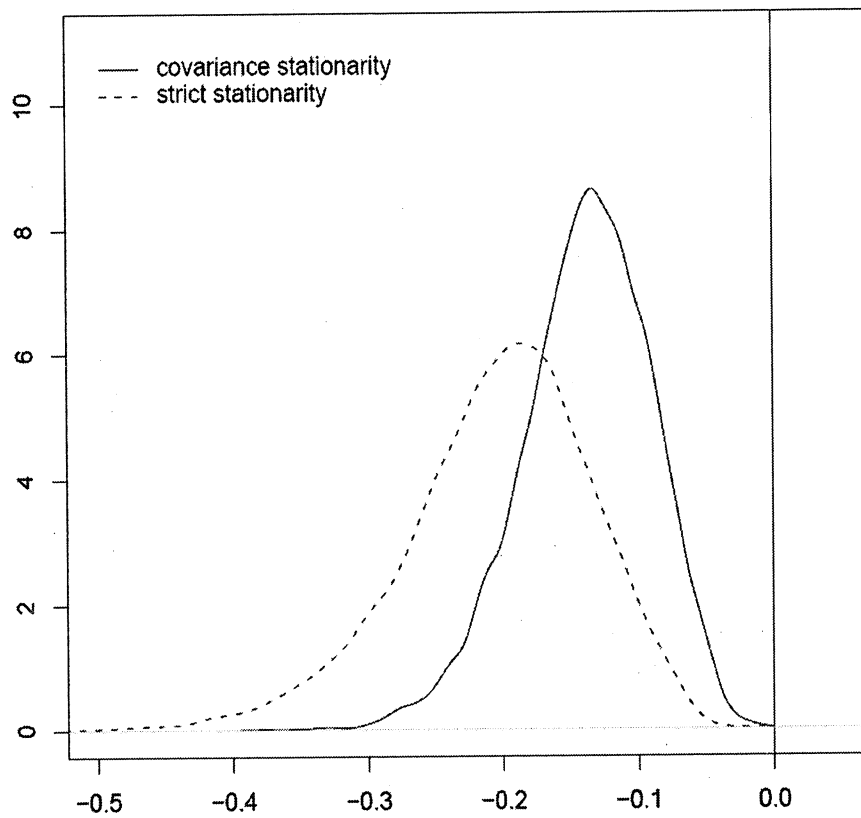


Figure 6. *CSC* and *SSC* posterior distributions. Gaussian kernel density estimates with bandwidth selected by the 'Silverman's rule of thumb' criterion (Silverman [12, page 48]).

References

- [1] Andrews, D. W. K., Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica*, 59, 3, May, 1991, 817-858.
- [2] Andrews, D. W. K. and Monahan, J. C., An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator, *Econometrica*, 60, 4, July, 1992, 953-966.
- [3] Chib, Siddhartha and Greenberg, Edward, Bayes Inference in Regression Models with ARMA(p,q) Errors, *Journal of Econometrics*, 64, 1994, 183-206.
- [4] Engle, Robert F., Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, *Econometrica*, 50, 4, 1982, 987-1008.
- [5] Gelman, Andrew and Rubin, Donald B., Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, 7, 4, 1992, 457-472.
- [6] Geweke, John F., Bayesian Inference and Posterior Simulators, *Canadian Journal of Agricultural Economics*, 49, 2001, 313-325.
- [7] Geweke, John F., Getting it Right: Joint Distribution Tests of Posterior Simulators, *Journal of the American Statistical Association*, 99, 467, September, 799-804, 2004.
- [8] Geweke, John F., Simulation Methods for Model Criticism and Robustness Analysis, in *Bayesian Statistics*, 6, 275-299, 1999, in Berger, J. O. and Bernardo, J. M. and Dawid, A. P. and Smith, A. F. M., Oxford: Oxford University Press.
- [9] Nakatsuma, Teruo, A Markov-Chain Sampling Algorithm for GARCH Models, *Studies in Nonlinear Dynamics and Econometrics*, 3, 2, 1998, 107-117.
- [10] Nelson, Daniel B., Conditional Heteroskedasticity in Asset Returns: A New Approach, *Econometrica*, 59, 2, 1991, 347-370.
- [11] Phillips, P. C. B. and Perron, Pierre, Testing for a Unit Root in Time Series Regression, *Biometrika*, 75, 1988, 335-346.
- [12] Silverman, B. W., *Density estimation for statistics and data analysis*, London: Chapman and Hall, 1986, First.

Mechanical Representations of Oja's Spatial Median

Richard William Farebrother

1 Castle Road, Bayston Hill

Shrewsbury, United Kingdom.

Email: R.W.Farebrother@ Manchester.ac.uk

Abstract : Farebrother (1987, 2002) has shown that two mechanical models based on strings and pulleys can be constructed for the least absolute deviations line fitting problem. In this paper we show that these mechanical models can easily be adapted for use with Oja's (1983) spatial median.

Key words: Duality of points and lines; L_1 -norm; least absolute deviations; linear programming; mediancentre; models based on strings and pulleys.

1 Objective Function

The history of geometrical and mechanical models of the mediancentre or centre of population has been outlined by Kuhn (1967) and Franksen and Grattan-Guinness (1989). In this paper we seek to generalise Farebrother's (1987, 2002) mechanical models for the least absolute deviations fitting of a straight line to a generalisation of the mediancentre proposed by Oja (1983).

It is well-known, see Farebrother (1992), that a triangle with vertices at (x_i, y_i) , (x_j, y_j) and (x_0, y_0) has an area of $\frac{1}{2}|d_{ij}|$, where

$$d_{ij} = \det \begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_0 & y_0 \end{bmatrix}$$

$$= (x_j - x_i)y_0 + (x_i y_j - x_j y_i) - (y_j - y_i)x_0$$

Oja's (1983) spatial median is defined by the values of x_0 and y_0 which minimise the sum of the areas of all $m = n(n-1)/2$ such triangles, $f = \sum_{i < j} |d_{ij}|$. Further, Niinimaa, Oja and Nyblom's (1992) fitting procedure takes the form of a

linear programming problem that seeks to minimise this function over all values of x_0 and y_0 .

2 First Mechanical Model

Simple mechanical models for Oja's spatial median may readily be obtained provided that the data have been rotated in such a way that all the x -values are distinct. In this context, the objective function becomes

$$f = \sum_{i < j} w_{ij} |e_{ij}|$$

where

$$w_{ij} = |x_j - x_i| e_{ij} = y_0 - a_{ij} - x_0 b_{ij} \quad a_{ij} = (x_j y_i - x_i y_j) / (x_j - x_i) b_{ij} = \frac{y_j - y_i}{x_j - x_i}$$

In this context, the problem takes the form of the weighted least absolute deviations fitting of a line of the form

$$a = y_0 - b x_0$$

to a grid of m points $(b_{ij}, a_{ij}) \quad i < j = 1, 2, \dots, n$ in the ba -plane.

Following Farebrother (2002, pp.101–106), we associate a real horizontal plane with the abstract ba -plane, then at each point (b_{ij}, a_{ij}) in this plane we drill a hole. Through this hole we pass a length of string with a weight w_{ij} at the lower end and a small ring at the upper end. We now place a rigid rod at an arbitrary position in the ba -plane and pass the m small rings over this rod in such a way that the associated strings run parallel to the a -axis from the ring to the hole. When this rod is in equilibrium, it will identify the values of x_0 and y_0 defining Oja's spatial median.

3 Second Mechanical Model

Unfortunately, this natural representation of the fitting problem has the defect that it lies in ba -space. To obtain the corresponding representation of the problem in xy -space, we determine the altitude of the ij th triangle $|d_{ij}|/c_{ij}$ by dividing twice the area of the triangle $|d_{ij}|$ by the length of its base

$$c_{ij} = \sqrt{[(x_i - x_j)^2 + (y_i - y_j)^2]}.$$

Thus we have to replace the m line segments connecting the m pairs of points by rigid rods of arbitrary length, pass a length of string over the ij th rigid rod and attach a weight proportional to the length of the ij th line segment c_{ij} to the lower end of the ij th string and connect the upper ends to a ring in the xy -plane. When in equilibrium

the strings will lie at right angles to the associated rods and the position of the ring will again identify the values of x_0 and y_0 defining Oja spatial median.

To aid a comparison of this second model with that of Farebrother (2002, pp.111–113), we note that the arbitrary point (x_0, y_0) is at distance e_{ij} from the fitted ij th line when distances are measured parallel to the y -axis and at distance $|d_{ij}|/c_{ij} = e_{ij}/\sqrt{1 + b_{ij}^2}$ when distances are measured perpendicular to the individual line segments.

4 Concluding Remarks

The mechanical models described in this paper may readily be adjusted to the q -dimensional generalisation of Oja's spatial median (with $q \geq 2$) provided that all $q \times q$ submatrices of the first q columns of the $n \times (q + 1)$ data matrix are nonsingular. For an alternative mechanical model based on elastic soap films, see Farebrother (2002, pp.58–59).

References

- [1] Farebrother, R. W. (1987), Mechanical representations of the L_1 and L_2 estimation problems, in Y. Dodge (Ed.), *Statistical Data Analysis Based on the L_1 -Norm and Related Methods*, North-Holland Publishing Company, Amsterdam, 455–464.
- [2] Farebrother, R. W. (1992), Geometrical foundations of a class of estimators which minimise sums of Euclidean distances and related quantities, in Y. Dodge (Ed.), *L_1 -Statistical Analysis and Related Methods*, North-Holland Publishing Company, Amsterdam, 337–349.
- [3] Farebrother, R. W. (2002), *Visualizing Statistical Models and Concepts*, Marcel Dekker Inc., New York.
- [4] Franksen, O. I. and I. Grattan-Guinness (1989), The earliest contribution to location theory? Spatio-economic equilibrium with Lamé and Clapeyron 1829, *Mathematics and Computers in Simulation* **31**: 195–220.
- [5] Kuhn, H. W. (1967), On a pair of dual nonlinear programs, in J. Abadie (Ed.), *Methods of Nonlinear Programming*, North-Holland Publishing Company, Amsterdam, 29–54.
- [6] Niinimaa, A., H. Oja. and J. Nyblom, (1992), The Oja Bivariate Median, *Applied Statistics*, **41**: 611–617.
- [7] Oja, H. (1983), Descriptive statistics for multivariate distributions, *Statistics and Probability Letters* **1**: 327–332.