Volume 48-50: 1 2016

Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 1, pp. 1-9

CRISIS IN SCIENCE? OR CRISIS IN STATISTICS! MIXED MESSAGES IN STATISTICS WITH IMPACT ON SCIENCE

D. A. S. FRASER AND N. REID
Department of Statistical Sciences, University of Toronto, Toronto, Canada M5S 3G3, Canada
Email: dfraser@utstat.toronto.edu, reid@utstat.toronto.edu

SUMMARY

Gelman and Loken (2014) draw attention to a "statistical crisis in science" and describe how risks with multiple p-values can be present even in the analysis of a single data set. There is indeed a crisis, as p-values are everywhere, in science, engineering, medicine, social science, health care, and the media; and conflicting results are misrepresenting the importance of p-values, and indeed of many disciplines themselves. We argue that risks of misinterpretation are widespread, but that the crisis is really in the discipline of statistics, and starts with mixed messages about the meaning and usage of p-values. These mixed messages then have downstream effects that seriously misinform scientific endeavours. What are these mixed messages concerning p-values? And should statistics continue with such messages that compromise the discipline? We discuss this and offer recommendations.

Keywords and phrases: Bayesian, frequentist, likelihood, multiple testing, p-value-function. significance function

AMS Classification: 62A01, 62F99

1 Introduction

This article is a response to Gelman and Loken (2014), who drew attention to a "statistical crisis in science" and showed how multiple p-values can arise, in good faith, in the analysis of a single data set. At about the same time, the *Journal of Basic and Applied Social Psychology* made headlines in *Nature* (Woolston, 2015) by deciding to no longer publish papers containing p-values. This debate continues, and there were several media reviews of news in December 2015 from CERN's Large Hadron Collider about a possible discovery of a new particle, and the associated "5-sigma" criterion commonly applied in high-energy physics (Castelvecchi, 2015; Spiegelhalter, 2015).

There is a crisis as p-values are everywhere, in science, engineering, medicine, social science, health care, and in the standard media phrase "19 times out of 20" commonly appearing in the reporting of polls. Our view is that while the risks of misinterpretation of p-values are widespread, the crisis is really in the discipline of statistics, in providing mixed messages about the meaning of a p-value. These mixed messages have downstream effects that can seriously affect all applications. We discuss this and offer recommendations.

2 Multiple meanings

2.1 The p-value function

Suppose we observe a variable, say y, that measures an unknown θ of interest; thus y is accessible through measurement, but θ is only indirectly accessible, through inference from y. If we had unlimited time and resources we could collect a great many values of the variable y and obtain the probability distribution of the variable y. This density indicates the stochastic behaviour of the variable, and if we assume that the form of the density is known, but its location (for example) is not, by identifying this via an unknown parameter θ we can view learning where the distribution is located as learning the value of θ. This could be, and often is, formalized by having a hypothesis, called a null hypothesis and designated H_0, that the unknown true value θ is θ_0; an example is indicated in Figure 1.

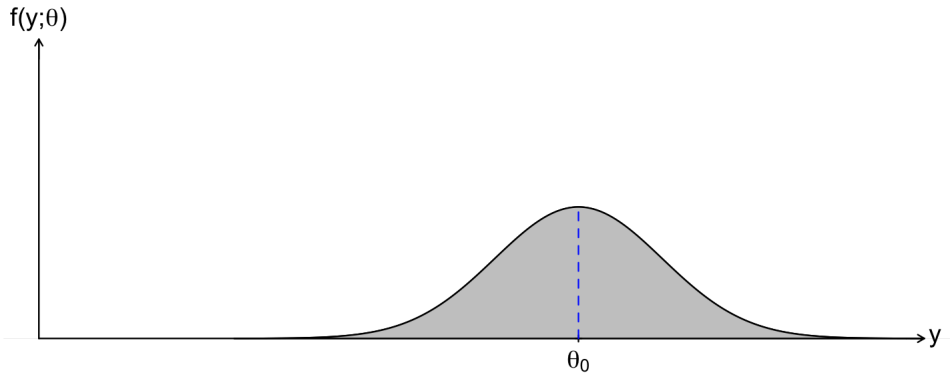

Figure 1: An accumulation of observations of y when the null hypothesis $H_0 : \theta = \theta_0$ holds.

Given a single observed measurement y^0, an investigator could then construct Figure 2, which shows that a proportion 6.1% of the distribution $\theta = \theta_0$ falls to the left of the observed measurement y^0, and 93.9% falls to the right. The observed p-value associated with H_0 would then be $p^0 = 6.1\%$ and is thus presenting just the percentile or statistical position of the data y^0 under H_0, or recording just a pure statement of factual information. As a definition this aligns with Fisher's 1920 proposal, later clarified in Fisher (1956).

This example is simplified to an extreme, but asymptotic arguments developed in Fraser (1990), Fraser and Reid (1993), and Brazzale et al. (2007, Ch. 8) show in wide generality that there is in fact such an approximating location model relevant to a single parameter of interest and that it can can be calculated quite routinely with more complex and realistic models.

Common statistical custom and usage don't usually stop with this percentile position, but proceed from the statistical position to scientific statements with potentially huge impact. For example, in high-energy physics, θ_0 could represent the mean value under background radiation, and then

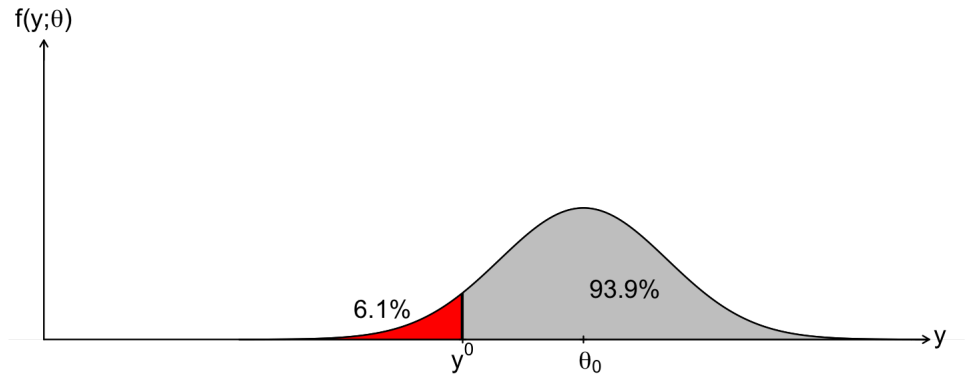

Figure 2: An observed data point y^0 and proportions left and right of the data under the hypothesis H_0.

larger values $\theta > \theta_0$ could indicate a new particle, such as the Higgs boson. In what sense does the p-value provide support for this new particle?

More informative than a single p-value, the p-value function $p(\theta)$, records the statistical position of the observed data y^0 for a range of values of θ: see Figure 3. This function presents the "statistical position" of the observation. It does not single out particular alternatives to θ_0, but leaves this choice to appropriate judgement in an application context Fraser (2014).

The p-value function presents in one plot all possible confidence bounds: we could for example solve $0.95 = p(\theta; y^0)$, the solution of which, $\hat{\theta}_L$ say, is a lower confidence bound at the conventional 95% limit. Under repeated observation of y from the model, the interval $(\hat{\theta}_L, \infty)$ will include the true value of θ 19 times out of 20, on average. The p-value function has also been called the confidence distribution function, e.g. in Cox (1958), Efron (1993), Xie and Singh (2013), Hjort and Schweder (2016). The p-value function or confidence distribution function has the added benefit that the direction of departure is recorded, as well as the magnitude.

2.2 Decision theory

Calculating observed proportions such as 0.061 and 0.939 as above was historically often challenging, and reference values corresponding to one or several standard values such as 5%, 10%, 90%, and 95% were derived and recorded in tables. Then in an investigation a statement such as "significant at the 10%" level, or "not significant at the 5% level", would be offered for the data point y^0 in Figure 2.

With the development of the theory of hypothesis testing by Neyman and Pearson (1933), this practice acquired a formal theoretical status. In due course the original concept of a p-value or observed level of significance as the position of the data with respect to the model changed its presentation into a decision for or against the hypothesis H_0, at some chosen fixed level of significance. The observed value y^0 then became a decision for, or against, some null value. Taking such de-

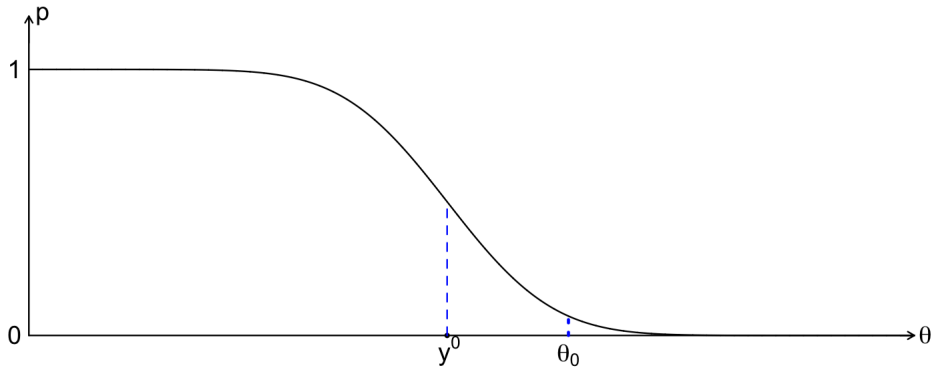

Figure 3: The observed p-value function from data y^0 as in Figure 3; height at θ_0 is 6.1%.

cisions at face value is a substantial change from the notion of statistical position, and has had the profound and unfortunate consequence of setting an arbitrary standard for determining the adequacy, or even publishibility, of the results from an experiment.

When p-values are used only to make a decision, and a larger sample size is viewed as a route to getting to the decision point faster, the results can be even more misleading.

Gelman and Loken express this concern for treating p-values from a decision theoretic viewpoint: "By convention, a p-value below 0.05 is considered a meaningful refutation of the null hypothesis: however, such conclusions are less solid than they appear". They do not, however, dwell further on this point. Many contemporary presentations of introductory statistics also overlook such concerns.

The point was emphasized famously in Ioannidis (2005), but there is a much earlier literature warning about this. Sterling (1959) wrote of "publication decisions and their possible effects on inferences drawn from tests of significance"; in particular "... (where) a borderline between acceptance and rejection is taken (at a) fixed point (say) 0.05 ... is interesting by itself ... (and when) used as a critical criterion for selecting reports for (publication) in professional journals (might result in) unanticipated results." Rozeboom (1960) wrote of "The fallacy of the null-hypothesis significance test" and quoted a famous philosophical epigram that the "accept-reject" paradigm is the "glory of science and the scandal of philosophy", meaning the glory of statistics and the scandal of logic and application.

2.3 Bayesian view of p-value

To this point we have assumed that the model for Figure 1 provides the full background information for θ. Another approach is available if we have a function $\pi(\theta)$ allegedly describing a probability density for potential values of θ. If the joint model is then accepted as valid, the application of the basic rules of conditional probability enable calculation of a probability distribution for θ, given the

observed measurement y^0, as $f(\theta \mid y^0) = c\pi(\theta)f(y;\theta)$. We can then compute, for example, the probability that θ is larger than θ_0, having observed y^0.

But where does such a probability density function $\pi(\theta)$ come from? Efron (2013) cites two possibilities: there may indeed be a case in some applications where randomness for the source of the true θ can be identified with a distribution $\pi(\theta)$: he calls this a genuine prior. If θ represents the rate of defectives in a manufacturing process, there may be enough data from previous manufacturing runs to identify such a distribution.

An alternative construction of a distribution $\pi(\theta)$ is by describing symmetries among various θ values: Efron (2013) calls these Laplace priors, as they received special support from Laplace (1812). In that case the construction of $f(\theta \mid y^0)$ can be regarded as a completely formal exercise, not embodying any probability interpretation. In this setting the best we could argue is that these probabilities have a meaning in as much as they lead to identical conclusions as the p-value function. Then the probability interpretation of the result is vacuous, but not misleading.

This is the case, for example, in a simple location model with a uniform prior for θ. The frequency calculation and the Bayes posterior probability calculation are computational reflections of each other; thus $s^0(\theta_0) = \int_{\theta_0} f(\theta \mid y^0)d\theta$ attaches the same value, 6.1%, to the statement that θ is larger than θ_0 as the argument above attaches to the probability under the model $f(y;\theta_0)$ that y is less than y^0: the Bayes posterior bound is in fact exactly a confidence bound: see Figure 4.

In our view the two ingredients $\pi(\theta)$ and $f(y;\theta)$, even if $\pi(\theta)$ is a genuine prior, should be left separate, rather than being combined into a joint model $\pi(\theta)f(y;\theta)$ describing the pair (y,θ). This makes available the full background information, and leaves to the concerned user the option to combine them if desired. This point is discussed from a slightly different point of view in Cox (2006, Ch. 5) and Cox and Reid (2015),where it is argued that "personalistic" priors have a different logical status from probability density functions.

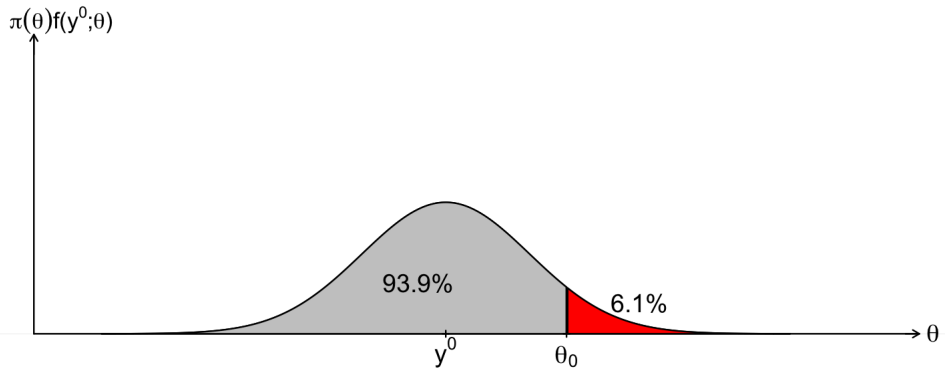

Figure 4: The Bayes calculation with the Laplace noninformative prior can with location symmetry duplicate the frequency calculation, thus giving a confidence result.

3 Responsibility and Risks

We have discussed three different interpretations for "p-value" or "level of significance": (i) The frequency view: The statistical position of the observed data with respect to a θ_0 value being tested; (ii) The decision theory view: The conventional level at which the data is just significant with respect to a θ_0 value being tested; and (iii) The Bayes view: The Bayes survivor calculation at a θ_0 value using some prior distribution for θ.

It is our view that the discipline of statistics should acknowledge responsibility for the consequences of the confusion, in many areas of application, caused by these multiple meanings. Fraser (2014) highlights three historically prominent cases where responsibility for statistical steps seems overwhelming, even in legal senses. The launch of the space-shuttle Challenger failed on January 28, 1986, causing seven deaths: statistical data available before the flight indicated a concern with the effect of low temperatures on critical O-rings, but the statistical warnings were by-passed (Dalal et al., 1989). The pain relief drug Vioxx was approved by the US Food and Drug Administration in 1999, but withdrawn by the pharmaceutical company in 2004 after evidence for an elevated risk of heart attacks became overwhelming, although statistical assessments as early as 2000 had indicated heightened risk of such serious events (Abraham, 2005). An estimated 40,000 people died and a five billion dollar settlement with the pharmaceutical company was obtained for those injured or the survivors (ONeil, 2012). Before the L'Aquila earthquake on April 5, 2009 an official committee with statistical expertise underemphasized in public statements the risk of an imminent major earthquake; some 300 died in that earthquake, and seven committee members were convicted of manslaughter (Marshall, 2012; Prats, 2012), a conviction that was overturned on appeal for six of the members (Abbott and Nosengo, 2014).

These examples emphasize that a misleading use of statistics can have serious consequences in lives lost and in billions of dollars in costs. These consequences can start with conflicting messages from statistics, and in particular the message that "statistical significance" is treated as an absolute, a decision, and that the goal of the statistical analysis of an observed set of data is to reach that elusive bar: a theme very common to applied work, especially among those new to the research process.

Gelman and Loken (2014) focus their discussion on the decision theoretic interpretation and address the consequences from this approach, emphasizing in particular the problem that for a given scientific or social scientific problem, the translation of "interesting science" to "statistical hypothesis" can, and often does, involve several hypotheses, and hence the calculation of several p-values, with a particular data set. They write "It would take a highly unscrupulous researcher to perform test after test in a search for statistical significance ... at the 0.05 level ... The difficult challenge lies elsewhere". They further note "it is reasonable for scientists to refine their hypotheses in light of the data". Their assessment of the risks emphasizes that the formulation of an hypothesis in science or social science is not as straightforward as identifying a single θ_0, and as a result multiple testing is implicit in a great many analyses, and more subtle than carrying out several tests in search of "$p < 0.05$".

We agree with them that the risks of using arbitrary p-values to define 'significance', and using these as decisions is very serious when multiple formulations of hypotheses lead explicitly or implicitly to large numbers of p-values. Among their recommended strategies of pre-registration,

authentic replication, and analysis of "all data", they include a claim "that p-values should not necessarily be taken at face value". This last we disagree with! It is the conventional but unwarranted attribution of decision, and the use of p-values for journal management, that are at the heart of the problem.

The p-value and p-value function is simply recording the statistical position of data relative to an hypothesis; it is elemental and provides an appropriate starting point for inference conclusions. It can guide the judgments about scientific conclusions, but cannot replace them. The consensus judgment in high-energy physics is that a 'discovery' is claimed when the p-value is less than 1 in 3.5 million: it is called "5-sigma" as this is the probability that a normal variable is greater than five standard deviations from the mean, the normal here being an approximation to the Poisson count of number of observed particles. Another physics example that received wide publicity in the popular media of the time was Eddington's verification of Einstein's theory of general relativity. The orbit of Mercury had been known in the 18-hundreds to precess at a rate different from that predicted by Newtonian mechanics, and Einstein's general relativity provided an adequate explanation. But further corroboration seemed appropriate to the physics community. General relativity also predicts the bending of light rays as they pass near a large mass; this provided, then, an appropriate variable to measure, and in May 1919 Eddington was able to carefully measure the apparent position of stars in the sky as indicted by light from the stars after it had passed adjacent to the sun during a solar eclipse.

Suppose, as viewed, the star light was passing on the right side of the hidden sun where general relativity would indicate that its apparent position in the sky was displaced to the right. Then if a 5-sigma event had been observed, the statistical position of the observed data would have been $p = 0.999,999,7$, indicating the large departure to the right; this value is the complement of $0.000,000,3$, in turn the reciprocal of 1 in 3.5 million. This p value records that data value was large, near 1; it is in the right tail of the null distribution under the standard theory of the time. The statistical position version of the p-value is appropriate and indicates the magnitude of the departure as well as the type of departure.

We believe the discussion is more urgent now, in the era of Big Data. As a reviewer has emphasized, the use of false discovery rates has been developed as a method of protecting against multiple hypothesis tests. In applications of many similar tests to a single set of data, for example in genome-wide association studies, this has provided some protection against claims of discoveries that could not subsequently be validated. Indeed the conventional, if somewhat arbitrary, 5-sigma rule of high energy physics is an *ad hoc* correction for multiple testing to protect exactly against false discoveries. This seems not to solve the issue, but rather to move the decision boundary.

An approach more directly aligned with the presentation of the p-value function is a method to correctly combine many such functions into a single summary p-value function. Methods of combination motivated by developments in the theory of composite likelihood are in development (Fraser and Reid, 2016).

For a great many settings where Big Data is available for analyses, the calculation of the dimensionality for possible hypotheses may be difficult or impossible, and the potential for making incorrect decisions is enormous. Attributing significance or decision to a comparison selected from

among millions of potential hypotheses suggests serious rethinking of the exploration process, the evaluation process, and the decision process. The risks for misleading decisions seem large; we could have mega p-values, mega decisions and mega wrong 'answers'. Scientists and social scientists are making serious efforts to address these issues; see for example the *Science* editorial McNutt (2014), and Gelman and Loken (2014)'s suggestions around pre-registration. Perhaps Statistics should stand up for its responsibilities before a Big Data Disaster.

4 Acknowledgement

We gratefully acknowledge discussions with Ian Spence in the Department of Psychology at the University of Toronto. This research has received support from the National Science and Engineering Research Council of Canada and the Senior Scholars Funding of York University. We thank reviewers of an earlier version for helpful suggestions for improvement.

References

[1] Abbott, A. and Nosengo, N. (2014). Re: Acquittal of 6 of the members of the committee. *Nature*, 515:7526.

[2] Abraham, C. (2005). Study finds Vioxx took deadly toll. *The Globe and Mail* 25 January 2005.

[3] Brazzale, A. R., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics*. Cambridge University Press, Cambridge.

[4] Castelvecchi, D. (2015). LHC sees hint of boson heavier than Higgs. *Nature News* 15 December 2015.

[5] Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29:357–372.

[6] Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press: Cambridge.

[7] Cox, D. R. and Reid, N. (2015). On some principles of statistical inference. *Intern. Statist. Rev.*, 83:293–308.

[8] Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-challenger prediction of failure. *J. Am. Statist. Assoc.*, 84:945–957.

[9] Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80:3–26.

[10] Efron, B. (2013). Bayes' theorem in the 21st century. *Science*, 340:1177–1178.

[11] Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

[12] Fraser, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika*, 77:65–76.

[13] Fraser, D. A. S. (2014). Why does statistics have two theories? In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present and Future of Statistical Science*, pages 237–252. CRC Press., Florida.

[14] Fraser, D. A. S. and Reid, N. (1993). Third order asymptotic models: likelihood functions leading to accurate approximation to distribution functions. *Statistica Sinica*, 3:67–82.

[15] Fraser, D. A. S. and Reid, N. (2016). On combining likelihoods and p-values. unpublished;

[16] Gelman, A. and Loken, E. (2014). The statistical crisis in science. *Amer. Scientist*, 102:460–465.

[17] Hjort, N. L. and Schweder, T. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press: Cambridge.

[18] Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2. e124. doi:10.1371/journal.pmed.0020124.

[19] Laplace, P. S. d. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier.

[20] Marshall, M. (2012). Seismologists found guilty of manslaughter. *New Scientist*, 22 October 2012.

[21] McNutt, M. (2014). Journals unite for reproducibility. *Science*, 346:679.

[22] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of a statistical hypothesis. *Phil. Trans. Roy. Soc. A*, 231:289–337. Reprinted in *Joint Statistical Papers of J. Neyman and E.S. Pearson*, Cambridge University Press, Cambridge, 1967.

[23] O'Neil, C. (2012). How Big Pharma cooks data – the case of Vioxx and heart disease. `http://www.nakedcapitalism.com/2012/02/25244.html`.

[24] Prats, J. (2012). The L'Aquila earthquake: Science or risk on trial? *Significance*, 9:13–16.

[25] Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57:416–428.

[26] Spiegelhalter, D. (2015). What are sigma levels? *Plus Magazine*, 18 December 2015.

[27] Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *J. Amer. Statist Assoc.*, 54:30–34.

[28] Woolston, C. (2015). Psychology journal bans P values. *Nature News*.

[29] Xie, M. G. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review (with discussion). *Intern. Statist. Rev.*, 81:3–39.

Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 1, pp. 11-12

COMMENTARY ON "CRISIS IN SCIENCE? OR CRISIS IN STATISTICS! MIXED MESSAGES IN STATISTICS WITH IMPACT ON SCIENCE"

A. GELMAN
Department of Statistics and Department of Political Science
Columbia University, New York, NY 10027, USA
Email: gelman@stat.columbia.edu

I agree with Fraser and Reid that the many abuses of p-values in the real world arise not so much because applied researchers have ignored the lessons of statistics, but because in many ways they have learned the lessons of statistics all too well. Thus, to the extent that education from statisticians to practitioners is part of the solution, what is needed is not merely to shout existing messages even louder, or to require scientists to take more of the usual sort of statistics courses, but rather for we, the statistics profession, to think carefully about what messages we are sending and how these messages can be encouraging statistics abuse.

Before going on, let me clarify that I think the real problem is not with p-values but with what is called "null hypothesis significance testing" (NHST): the practice by which a researcher seeks to reject a straw-man null hypothesis as evidence in favor of some favored alternative. This is rife in the literature, indeed is the standard use of statistical analysis in psychology, medicine, and other fields. Heres an example, one of many: Carney, Cuddy, and Yap (2010) presented evidence that when subjects held their body in a certain posture called the power pose, they gained a feeling of confidence and certain hormone levels increased, compared to a control position in which the subjects held an alternative posture. Key comparisons in this paper had p-values of less than 0.05. A few years later, Ranehill et al. (2014) did a larger-scale preregistered replication of the study and failed to find an effect. What went wrong? The problem was with the logic of the significance test: In their original paper, Carney, Cuddy, and Yap had the choice of many possible data analyses; as a result, the probability of attaining statistical significance in some way would be much higher than 5%, even in the absence of any effect. Indeed, effects are small enough and variation is high enough that it would be essentially impossible to untangle signal from noise in a study of that size. The problem with the p-value here is that it is contingent on the choice of what analyses might have been performed, had the data been different.

In the Cuddy, Carney, and Yap study, a similar problem would have arisen with the NHST logic even had some other method than p-values been used. For example if likelihood ratios or Bayes factors were used to determine statistical significance and were used to reject the null, one would again have to be concerned about the many forking paths in this analysis.

Fraser and Reid discuss a Bayesian interpretation of the p-value that is similar to that presented by Greenland and Poole (2013). In my discussion to that Greenland and Poole paper (Gelman,

2013), I wrote that I see the mathematical connection between the (one-tailed) p-value and the posterior probability $\Pr(\theta > 0 \mid y)$ under a uniform prior distribution on θ — but I question that uniform prior. The problems where NHST is causing the most problems are where effects are small. For example, Gertler et al. presented a study of early childhood intervention in Jamaica, reporting an effect of 42% on adult income. (It was a longitudinal study in which the children, first observed before school age, were followed up into their twenties.) The estimate was statistically significant; thus the 95% confidence interval on the treatment effect was something like [2%, 82%]. But I don't believe this interval. I certainly don't believe the 82% on the high end and, thinking Bayesianly, my prior based on the literature of such interventions is that any effect will be small. Perhaps a normal prior with mean 0 and standard deviation 10% would be reasonable, in which case the resulting posterior inference would not nearly be so optimistic as implied by the uniform prior. From a frequentist perspective, I do not think this interval has good coverage because of selection—"researcher degrees of freedom", in the words of Simmons, Nelson, and Simonsohn (2011)—in the data processing and analysis.

The problem that I see in statistical education is that we present statistical methods as alchemy, a way to convert randomness into a sort of certainty, as associated with words such as "confidence" and "significance". Look at statistics textbooks — including my own! — and you'll see example after example in which data are collected, analysis is done, and then inference is conveniently summarized with statistically significant p-values and confidence intervals that comfortably exclude zero. It's no wonder that practitioners, trained from such books, go out into the world expecting to find such clean summaries. The message we (implicitly) teach is that if you're studying a real effect and you have a good design and reasonable sample size, you'll succeed in the sense of getting a low p-value or a high posterior probability or a confidence interval that excludes zero.

Now consider this from the point of view of a researcher, Dr. X, analyzing some data. Dr. X presumably thinks hes studying a real effect (otherwise why work on the problem at all) and that he did a good design, and he might have even performed a power analysis to check that his sample size is large enough. Such power analyses are typically wildly optimistic because published effect size estimates tend to be way too large, biased as they are by the statistical significance filter: big estimates are statistically significant and get published, while estimates near zero, being non-significant, never appear. But this is a subtle point not mentioned in textbooks and not, we suspect, recognized by most researchers. So here is Dr. X, sure he's doing everything right and expecting to see a positive result: it's no wonder that he might jiggle his data a bit to get everything to line up.

So, to get researchers to stop chasing their tails with NHST, I think we need to revise our education, to take away the message that statistical significance (or the Bayesian or confidence interval equivalent) will come as a matter of course. Rather, researchers need to learn to live with uncertainty.

References

[1] Gelman A. (2012). P-values and statistical practice. *Epidemiology* **24**(1), 69–72

[2] Greenland S. and Poole C. (2012). Living with p values: resurrecting a Bayesian perspective on frequentist statistic. *Epidemiology* **24**(1), 62–68.

Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 1, pp. 13-13

REJOINDER: CRISIS IN SCIENCE? OR CRISIS IN STATISTICS...

D. A. S. FRASER AND N. REID
Department of Statistical Sciences, University of Toronto, Toronto, Canada M5S 3G3, Canada
Email: dfraser@utstat.toronto.edu, reid@utstat.toronto.edu

We'd like to thank Andrew Gelman for the thoughtful discussion of our note, and for the article that inspired our response. That paper (Gelman and Loken, 2014) expressed concerns for a crisis in science; our response argued that the crisis was in statistics, with its wide-spread recommendation that p-values be represented in terms of decisions, at the 5% level, or even the 5 sigma level or 1 in 3.5 million as recently used by High Energy Physicists.

The commentary agrees with our perspective on "NHST", and provides insightful examples from applications. Technical concerns aside, there are also issues of responsibility, professionalism, and ethical behaviour that can't be overlooked. It seems then that we are in full agreement on the substance of the issues, with some differences of opinion on how the concerns should be weighted.

May Statistics rise to its challenges.

References

[1] Gelman, A. and Loken, E. (2014). The statistical crisis in science. *Amer. Scientist*, 102:460–465.

Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 1, pp. 15-33

Analysis of ordinal longitudinal data using semi-parametric mixed models

KALYAN DAS
Department of Statistics, University of Calcutta, India
Email: kalyanstat@gmail.com

SURUPA ROY
Department of Statistics, St. Xavier's College, Kolkata, India
Email: surupachakraborty@yahoo.co.in

ASIS KUMAR CHATTOPADHYAY
Department of Statistics, University of Calcutta, India
Email: akcstat@gmail.com

SUMMARY

A spline mixed item response theory model that allows for three-level multivariate ordinal outcomes and accommodates multiple random subject effects is proposed for analysis of ordinal outcomes in longitudinal studies. Assuming cumulative logit model with proportional odds, maximum marginal likelihood estimation for model parameters is proposed utilizing Monte Carlo Metropolis Hastings Newton Raphson (MCMHNR) algorithm. An iterative Fisher scoring solution, which provides standard errors for all model parameters, is considered. The performance of the estimates of the model parameters in finite samples has been looked into. A longitudinal orthodontic data set, where plaque content in teeth is repeatedly measured over time, is used to illustrate application of the proposed model.

Keywords and phrases: ordinal response, proportional odds model, spline, Monte Carlo EM, Metropolis-Hastings, orthodontic data.

1 Introduction

Many interesting problems in Biomedical, industrial and other experiments involve the study of how an ordered response variable depends on a set of regressors. In psychometric and educational testing literature, a large amount of research has been devoted to developing mixed-effects models for subject-specific comparisons of multivariate ordinal responses. In longitudinal studies, information from the same set of subjects is measured repeatedly over time. Multivariate data arise when different item responses, related to a single underlying outcome, are measured to provide more complete and reliable information. The aim of such studies is to estimate the mean or individual response at a

certain time, to relate time-invariant or time-dependent covariates to repeatedly measured response variables, or to relate the response variables to each other.

One way to model ordinal regression data is to assume that the observed response is the discrete version of a continuous latent variable for which a linear regression model holds. Alternatively, an index model of the discrete probabilities may be written for a given transformation, called link function as in the seminal paper of McCullagh (1980). It is well known that the latent variable approach and the index model approach are essentially equivalent (see Greene, 2004 and Wooldridge, 2003). Examples of such related models are obtained by assuming the logistic distribution for the errors in the latent variable and the ordered logit model, or the normal distribution for the latent error and the ordered probit model.

The restricted version of the generalized logit model is the standard ordered logit model discussed in most statistics textbooks and it is known in the statistical literature as the proportional odds model (see McCullagh, 1980). Especially when the number of possible ordinal values is large, the model may require many more parameters than the simple ordered logit model. This may be justified for example when it is reasonable to assume that the threshold between adjacent categories depends on subjective judgments, as for instance in the analysis of the determinants of health status, happiness etc. As the ordered logit model may be seen as a properly constrained generalized logit model, the effect of covariates on threshold parameters may be tested by imposing appropriate linear constraints. When the dependence of threshold parameters on individual covariates is not justified by the nature of the response variable, the rejection of the proportional odds assumption should be taken as a warning that the latent model is not properly specified, like when, for instance, the distribution of the error is heteroscedastic or the covariate is not exogenous.

Often in longitudinal studies it is required to characterize the temporal trends exhibited by some real data. The mean trajectory appears to show curvature. In fact individual series shows more curvature. In a situation where the primary focus of the analysis is to relate disease progression at different time points to the subject's habit/nature, it is of practical interest to develop an appropriate method that truly incorporates the temporal patterns as well as the covariate information. Certainly, a less restrictive assumption on the time functions might be more desired than imposing some parametric assumptions, which might be incorrect.There has been a tremendous advancement in statistical research on non parametric function estimation. In many situations a semi parametric generalized partially linear mixed model (GPLMM) is considered for handling the covariate effects (time) non-parametrically. Such a model is essentially a compromise between the GLMM and a fully nonparametric model. This kind of model is popular in longitudinal studies such as human viral dynamics, pharmacokinetic analyses and studies of growth and decay. On the other hand the inclusion of a nonparametric covariate in an otherwise GLMM raises the high dimension problem. In order to avoid this, we consider a generalized partial ordinal longitudinal model (GPOLM) that can be viewed as a compromise between GLMM and a fully nonparametric model.

Considerable studies have been done on partially linear models (see Hardle et al., 2000). In order to analyze discrete outcomes, where the influential covariates and the outcome have definite functional relationship (monotone), it is natural to extend the model to a partial semi parametric Generalized linear model. Previously, Severini and Staniswalis (1994), Hardle et al. (1998) and Muller

(2001) have looked into the influential aspects of GPLM. Later Lin and Carroll (2001a), Wang et al. (2005) and He et al. (2005) have considered the GPLMM in the context of clustered/longitudinal data. Lin and Carroll (2001b) address that the conventional profile kernel-based approach is incapable of producing a $\sqrt{n}$ consistent estimator of the parameters unless the non-parametric function is under-smoothed or working independence is assumed for the GEE methodology. These limitations can be avoided if regression spline approximation is considered in GPLMM.To the best of our knowledge, no literature has yet been published for the analysis of GPOLM. Our attempt is to show that in the regression spline approximation under GPOLM, the spline approach results in the optimal rate of convergence for estimating the unknown function and the parameters of interest. The primary focus of our paper is to use a spline mixed regression model for analyzing ordinal longitudinal data. Such a model accommodates longitudinal dependence and subject specific variation in the data through random effects. We consider a data on oral hygiene where 220 individuals consisting of students and staff members of medical schools in and around the city of Kolkata were selected randomly irrespective of age, sex and oral hygiene status and their plaque scoring was recorded according to Turesky et al. (1970). The reduction in the thickness of plaque for subjects are usually recorded as belonging to four different categories, viz 'no reduction', 'slight reduction', 'moderate reduction' and 'vast reduction' (to a great extent). In addition, auxiliary information on age, sex, food habit, smoking habits etc were also observed for each subject. The purpose of the study is to see whether the progression of the plaque reduction is truly effective with the use of a solution (kept in mouth for 1 minute followed by a thorough rinse with water to remove any excess of disclosing solution) and if so, to what extent such progression depends on the covariates taken.

The article is organized as follows. In Section 2 we introduce the spline mixed cumulative logit model with proportional odds setup. In Section 3 we consider estimation of model parameters using MCMHNR approach. In section 4 an asymptotic study is given. An exact sample study has been carried out in Section 5, to see the performance of the estimator under the proposed approach. Data arising from an orthodontic study have been analyzed in Section 6. Finally, conclusion and discussion are made in Section 7.

2 The Model and Likelihood

Consider a trial involving n individuals in which each individual is to be examined at K assessment times. Let y_{ijk} denote the ordinal response that has $L+1$ distinct levels, $0, \ldots, L$ (say) for individual i within the cluster j at the assessment time $(k, i = 1, \ldots, n;\ j = 1, \ldots, r; k = 1, \ldots, K)$. This gives rise to a hierarchical data structure where the assessment times (level 1) are nested within the clusters (level 2) which in turn are nested within the individuals (level 3). Further suppose, associated with the ordinal response, x_{ijk} denote the covariate vector for individual i in cluster j at time k. The covariates are completely known and may be assumed to be fixed across the entire observation times. Let u_{ij} denote the subject and cluster specific random component vector corresponding to individual i in the cluster j. The random component reflects the unobserved heterogeneity in the data. Dummy variables are often used to represent categorical variables in estimation of parameters. Let us denote Y_{ijk} as a vector of $L+1$ indicator variables, given by, $Y_{ijk} = (Y^0_{ijk}, \ldots, Y^L_{ijk})'$ with $Y^l_{ijk} = 1$, if

$y_{ijk} = 1$ and 0, otherwise ($l = 0, \ldots, L$). Further suppose, the vector of probabilities and cumulative probabilities are respectively denoted by $\pi_{ijk} = (\pi^0_{ijk}, \ldots, \pi^L_{ijk})'$ and $\eta_{ijk} = (\eta^0_{ijk}, \ldots, \eta^L_{ijk})'$, where π^l_{ijk} and η^l_{ijk} are given by,

$$\pi^1_{ijk} = P\left(Y^1_{ijk} = 1 \mid x_{ijk}, u_{ij}\right) = P\left(y_{ijk} = l \mid x_{ijk}, u_{ij}\right), \tag{2.1}$$

$$\eta^1_{ijk} = P\left(y_{ijk} \leq l \mid x_{ijk}, u_{ij}\right) = \sum_{i0}^{1} \pi^1_{ijk}. \tag{2.2}$$

Corresponding to the individual i, the multivariate ordinal data can be represented as $(y_{i11} = c_{11}, \ldots, y_{ijk} = c_{jk}, \ldots, y_{irK} = c_{rK})'$, where c_{jk} $(j = 1, \ldots, r; k = 1, \ldots, K)$ can take the ordinal scores $0, \ldots, L$. Conditional on the subject and cluster specific random components u_{ij} and given the covariates, the associated probability follows from (2.1) and can be written as,

$$\begin{aligned} P_{ij} &= \prod_{k=1}^{K} \prod_{l=0}^{L} \{P(y_{ijk} \leq l \mid u_{ij}, x_{ijk}) - P(y_{ijk} \leq l-1 \mid u_{ij}, x_{ijk})\}^{l(y_{ijk}=l)} \\ &= \prod_{k=1}^{K} \prod_{l=0}^{L} (\eta'_{ijk} - \eta^{l-1}_{ijk})^{l(y_{ijk}=l)}, \end{aligned} \tag{2.3}$$

where $I(y_{ijk} = l) = 1$, if $y_{ijk} = l$ and 0 otherwise, $\eta^{-1}_{ijk} = 0$ and $\eta^L_{ijk} = 1$. To model the dependence of the response on the covariates and the random component we use cumulative logit model with proportional odds assumptions. Typically such a model is written as,

$$\text{log it}(\eta'_{ijk}) = \log\left(\frac{\eta'_{ijk}}{1 - \eta'_{ijk}}\right) = \lambda_l + x'_{ijk}\beta + z'_{ijk}u_{ij} + f_0(t_{ijk}), \tag{2.4}$$

where λ_l $(l = 0, \ldots, L-1)$ is the intercept in the lth logit model which satisfy the relationship $\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{L-1}$ and β denotes the p dimensional vector of covariate effects corresponding to x_{ijk}. The random component vector u_{ij} is a subject and cluster specific random effect of dimension q associated with the completely specified design vector z_{ijk}. For subject i in cluster j, we write the random component vector u_{ij} as, $u_{ij} = (u^1_{ij}, \ldots, u^q_{ij})'$ and assume that $u_{ij} \sim N_q(0, I_q)$. Let us further write $u_i = (u'_{i1}, \ldots, u'_{ir})'$, where

$$u_i \sim N_{rq}(0, I_q \otimes \Sigma), \quad \Sigma = \sigma^2[(\mathbf{1} - \rho)I_r + \rho\mathbf{1}\mathbf{1}']. \tag{2.5}$$

In model (2.5), σ^2 and ρ denotes the intra cluster variability and correlation coefficient respectively. They are treated as nuisance parameters and are estimated along with the other regression parameters. In model (4), t_{ijk} may be simply time or in general any time dependent covariate and $f_0(.)$ is an unknown smooth function.

We use the basis of cubic B-splines with q preselected knots to approximate the unspecified smooth function f_0 in which the rth knot corresponds to the $r/(q+1)$th sample quantile of the distinct values of $t_{ijk} (i = 1, \ldots, n; j = 1, \ldots, r; k = 1, \ldots, K)$. Let $B_1(t), \ldots, B_{q+4}(t)$ be the cubic B-spline basis for the space of cubic splines with q preselected knots. For details on computing

of B-splines and their mathematical properties we refer to Boor (2001). The cubic B-splines space includes a constant function, and the constant is given in the parametric component of the model (4), so to model $f_0()$ one of the $(q+4)$ B-spline basis functions needs to be dropped so that the resulting parameterization is of full rank. Any one of them can be dropped, but for convenience Li (2011) models f_0 as a linear combination of the first $q+3$ fixed-knot cubic B-spline basis functions. In this paper in order to approximate f_0 by a regression spline, we consider a set of knots on $[0,1]$ with $0 = s_0 < s_1 < \cdots < s_{k_n} = 1$ and generate $N = k_n + l$ normalized B-spline basis functions of degree $l+1$ that span the linear space. We then express $f_0(t) \approx v'(t)\gamma$, where, $v(t) = (B_1(t), \ldots, B_N(t))'$ is the vector of basis functions and $\gamma \in R^N$ is the spline coefficient vector. Let us denote the vector of parameters by $(\theta', \phi')'$ where, $\theta = (\lambda_0, \ldots, \lambda_{L1}, \beta', \gamma')'$ and $\phi' = (\sigma^2, \rho)'$. Then in view of (3) and (5), the likelihood for subject i can be written as,

$$L_i(\theta, \phi) = \int \prod_{j=1}^{r} \prod_{k=1}^{K} \prod_{l=1}^{L} \left[\eta'_{ijk} - \eta^{l-1}_{ijk}\right]^{I(y_{ijk}=l)} g(u_i) du_i, \tag{2.6}$$

where $g(u_i)$ denotes the density function of u_i given in (2.5). Here our primary focus lies in estimating and making inference on the parameter vector θ although the vector of nuisance parameter ϕ is also estimated in the study simultaneously.

The critical issue for getting a rigorous model selection criterion can be based on estimating the relative expected Kullback-Leibler $(K - L)$ information. Akaike (1973) found that the maximized log likelihood value was a biased estimate of $K - L$ information but this bias was approximately equal to 'p', the number of estimable parameters in the approximating model. Thus an approximately unbiased estimator of $K - L$ information for large samples and good models is given by Akaike's Information Criterion (AIC), where

$$\text{AIC} = 2 \log L(\hat{\theta}, \hat{\phi}) + 2p. \tag{2.7}$$

In (2.7) above, $(\hat{\theta}, \hat{\phi})$ is the maximum likelihood estimator of the parameter vector arising in model (2.6) and $L(\cdot)$ denotes the likelihood function given the data vector. Minimizing the AIC over a set of possible models can thus be seen as minimizing the average distance of an approximating model to the underlying truth.

3 Parameter Estimation

The likelihood function given in (2.6) is difficult to maximize because of the multidimensional integral over u_i which is the consequence of a mixed effects modelling. Numerical integration techniques like Gauss Hermite quadrature or adaptive Gaussian quadrature (Pinheiro and Bates, 1995) can be used to approximate the above integral to any practical degree of accuracy. Diverse methodologies in both Bayesian and Classical paradigm are available in the literature for fitting GLMM. In Bayesian perspective Markov Chain Monte Carlo (MCMC) method is implemented via Gibbs sampling techniques (Zeger and Karim, 1991) to generate repeated samples from the posterior distribution of the random effects. In the classical approach Breslow and Clayton (1993)

proposed the penalized quasi likelihood (PQL) for approximating the high dimensional integration using Laplace approximation. However, as reported by several authors PQL estimates are biased downwards for some variance components. Later Breslow and Lin (1995) and Lin and Breslow (1996) gave bias corrected PQL. McCulloch (1994) investigated GLMM with a probit link using Monte Carlo EM (MCEM). He extended MCEM to the logit model and introduced the Monte Carlo Newton Raphson (MCNR) and simulated maximum likelihood methods. For simple models it was found that the MCNR estimates inherits the properties of the exact ML estimates. Natarajan et al. (2000) and Zhou and Liu (2008) used the Monte Carlo version of EM to calculate ML estimates of parameters. Meza et al. (2009) and Davier and Sinharay (2010) proposed an alternative to MCEM via the Stochastic Approximation EM (SAEM) of Deylon et al. (1999). We could have considered any one of the three stochastic versions (SEM, SAEM and MCEM) to analyze our data. Since all three lead to similar conclusions (Celeux et al., 1995), we preferred to work with MCEM method here.

In this paper we adopt the MCNR approach to calculate the fully parametric Maximum likelihood estimates based on the likelihood (6). The Monte Carlo approach calls for generating random observations from the posterior distribution of the random effects which however is not in a closed form. To circumvent this difficulty Metropolis Hastings algorithm (see Chib and Greenberg, 1995) is used to generate data from the posterior distribution of the random effects which does not require the exact form of the conditional distribution. Moreover a good starting solution is needed for the MCNR method. In our analysis moment estimates are used. McCulloch (1997) pointed out that although this approach is computationally intensive it provides feasible solutions for a variety of data configurations. In presence of influential points in the data this method can be extended to the Robust Monte Carlo Newton Raphson method of Sinha (2004).

3.1 The MCMHNR Approach

To set up the EM algorithm, we consider the random effects to be missing. We write the observed data for individual $i \;\; (i = 1, \ldots, n)$, as $D_{0i} = \{y_{ijk}, x_{ijk}, t_{ijk}; j = 1, \ldots, r;\, k = 1, \ldots, K\}$ and the complete data is denoted by $D_{ci} = \{y_{ijk},\; x_{ijk},\; t_{ijk};\, j = 1, \ldots, r, k = 1, \ldots, K\}$. Further suppose $f(D_{oi} \mid u_i)$ denotes the conditional distribution of the observed data given the random component. Then using (3) and (5) the complete data log likelihood for all the subjects is given by,

$$\begin{aligned} l_c(\theta, \phi) &= \sum_{i=1}^{n} \log(f(D_{0i} \mid u_i; \theta)) + \sum_{i=1}^{n} \log(g(u_i; \phi)) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{r} \sum_{k=1}^{K} \sum_{l=1}^{L} I(y_{ijk} = l) \log\left[\eta_{ijk}^{l} - \eta_{ijk}^{l1}\right] + \sum_{i=1}^{n} \log(g(u_i; \phi)) \\ &= l_{c1}(\theta) + l_{c2}(\phi). \end{aligned} \tag{3.1}$$

From (3.1) it is to be noted that since θ enters only the first term so the M step of EM algorithm with respect to θ uses only $L_{c1}(\theta)$. The second term in (8) involves only the distribution of u_i which is assumed to be normal and so maximizing the likelihood $l_{c2}(\phi)$ gives the standard maximum

likelihood estimates of ϕ after replacing u_i's with their conditional expected values. Writing $D_0 = \{D_{0i}, i = 1, \ldots, n\}$ and $u = (u_1', \ldots, u_n')'$, the score functions for θ and ϕ can be expressed as:

$$\xi_\theta(\theta) = E_u\left[\frac{\partial l_{c1}(\theta)}{\partial \theta}\Big| D_0\right] = 0; \quad \xi_\phi(\phi) = E_u\left[\frac{\partial l_{c2}(\phi)}{\partial \phi}\Big| D_0\right] = 0. \tag{3.2}$$

In order to solve for θ and ϕ from equation (3.2), we propose a Monte Carlo Newton Raphson (MCNR) algorithm. Using MCNR, the updated estimate of θ and ϕ at $(t+1)$th step is given by,

$$\theta^{(t+1)} = \theta^{(t)} - \Lambda_1^{-1(t)}\xi_\theta(\theta^{(t)}), \quad \phi^{(t+1)} = \phi^{(t)} - \Lambda_2^{-1(t)}\xi_\phi(\phi^{(t)}) \tag{3.3}$$

where $\Lambda_1^{(t)} = \partial\xi_\theta(\theta)/\partial\theta|_{\theta^{(t)}}$ and $\Lambda_2^{(t)} = \partial\xi_\theta(\phi)/\partial\phi|_{\phi^{(t)}}$. The expressions for first and second order derivatives are given in Appendix A1. The MCNR approach gives an iterative computational scheme, where the maximization step becomes automatic. However the conditional expectations in (3.2) cannot be computed in a closed form. This is because the conditional distribution of u involves the marginal distribution of the data which in fact is the likelihood in equation (2.6) that we are trying to avoid calculating directly. To circumvent this difficulty we use Metropolis Hastings algorithm (Smith and Roberts, 1993) to produce random draws from the conditional distribution of $u \mid D_0$. Then we can approximate the required expectation in (3.2) by Monte Carlo approach.

To implement the Metropolis algorithm, we first specify the candidate distribution $h(u)$ from which potential new values are drawn and then compute the acceptance function that gives the probability of accepting the new value (as opposed to keeping the previous value). In our case, the target density can be expressed as proportional to the product of the density $g(u;\phi)$ that can be sampled and the conditional density $f(D_0 \mid u, \theta)$ that is uniformly bounded. Thus following Chib and Greenberg (1995) we set the proposal density to be equal to $g(.)$ (as in the independence chain) to draw candidates. In this case the acceptance probability takes a simplified form and requires the computation of $f(D_0 \mid u, \theta)$ only. Let u^0 denote the previous draw and u^{can} is a new value from the candidate distribution. Then we accept u^{can} as a potential observation from the conditional distribution of with probability of acceptance given by,

$$A(u^0, u^{com}) = \min\left\{\frac{f(D_0 \mid u^{com}, \theta)}{f(D_0 \mid u^0, \theta)}, 1\right\}. \tag{3.4}$$

Incorporating the Metropolis step in MCNR method results in MCMHNR algorithm which can now be stated as follows:

Step 1: Choose starting values θ^0, ϕ^0. Set $t = 0$.

Step 2: Generate R values $u^{(1)}, u^{(2)}, \ldots, u^{(R)}$ from the conditional distribution $f(u \mid D_0, \theta, \phi)$ using the Metropolis Hastings algorithm and use them to form the Monte Carlo estimates of the expectations.

Step 3: Compute:

$$\theta^{(t+1)} = \theta^{(t)} - \hat{\Lambda}_1^{-1(t)}\hat{\xi}_\theta(\theta^{(t)})$$
$$\phi^{(t+1)} = \phi^{(t)} - \hat{\Lambda}_1^{-1(t)}\hat{\xi}_\phi(\phi^{(t)}).$$

Replacing the expectations in (3.2) by Monte Carlo estimates and using (3.1),it follows that,

$$\hat{\xi}_\theta(\theta) = \frac{1}{R}\sum_{r=1}^{R} \frac{\partial}{\partial\theta} \log f(D_0 \mid u^{(r)}; \theta); \hat{\xi}_\phi(\phi) = \frac{1}{R}\sum_{r=1}^{R} \frac{\partial}{\partial\phi} \log g(u^{(r)}; \phi)$$

$$\hat{\Lambda}_1 = \frac{\partial}{\partial\theta}\hat{\xi}_\theta(\theta); \hat{\Lambda}_2 = \frac{\partial}{\partial\phi}\hat{\xi}_\phi(\phi)$$

Set $t = t + 1$.

Step 4: If convergence is achieved, declare $\theta^{(i+1)}$ and $\phi^{(i+1)}$ as the maximum likelihood estimates of θ and ϕ respectively. Otherwise return to Step 2.

3.2 Knot Selection

An important aspect of spline smoothing is knot selection. Since we are mainly concerned with the efficiency of the covariate effect estimates, we opt for convenient choices of knot placements. For the Knot selection we have applied a data adaptive scheme which is briefed below:

Step 1: We at first consider $Q_1 = 10$ largest local maxima and $Q_2 = 10$ smallest local minima.

Step 2: We have identified the time points corresponding to these $Q = Q1 + Q2$ points. These Q points have been chosen as the initial knots. Let $q = Q + k + 1$, for cubic spline $k = 3$. These k points are determined based on the quantiles.

Step 3: We removed the ith knot and evaluated the residual sum of squares (RSS_i), for $i = 1, 2, \ldots$

Step 4: We have chosen that model for which RSS_i is minimum and set $q = q - 1$.

Step 5: We have continued Steps 2-4 till $q = k + 1$.

4 Asymptotics

In this section, to ensure consistency of the proposed estimates, the asymptotic properties of the solution to score equations in (3.2) have been investigated. The asymptotic distribution of the estimators of θ and ϕ would be separately looked into as in view of (3.1), l_{c_1} involves only θ and l_{c_2} involves only ϕ. Essentially, here this section, we would consider only the asymptotic distribution of $\hat{\theta}$ as that of $\hat{\phi}$ is straightforward. We consider a sequence of consistent estimators $\hat{\theta}_n (= \hat{\theta}$ say) in the sense that as

$$n \to \infty, \sup_{t\in[0,r]} \mid v'(t)\hat{\gamma} - f_0(t) \mid \xrightarrow{P} 0, \hat{\lambda} - \lambda^0 \xrightarrow{P} 0 \text{ and } \hat{\beta} - \beta^0 \xrightarrow{P} 0,$$

where λ^0 and β^0 are true unknown values of λ and β respectively. The required basic assumptions are given below.

A.1 The distinct values of $t_{ijk}, 0 \leq t_{ijk} \leq \tau$ form a quasi-uniform sequence that grows dense on $[0, 1]$.

A.2 For every $i, Max\{\|X_i\|\} \leq B_0$ for some non-random constant B_0, where $X_i = ((x_{ijk}))$ $i = 1, \ldots, n; j = 1, .., r; k = 1, \ldots, K$.

A.3 $\mid f_0^{(s)}(.) \mid < A_0$, for some non-random value A_0 for $s \geq 2$.

A.4 Conditional on data and for every $i, \sup_{i\geq 1} E\|S_{ic}\|^{2+\delta} < \infty$, for some $\delta > 0$, where $S_{ic} = \frac{\partial}{\partial\theta} l_{ic}(\theta)$ and $l_{ic}(\theta) = \sum_{i=1}^{r} \sum_{k=1}^{K} \sum_{l=1}^{L} I(y_{ijk} = l) \log[\eta'_{ijk} - \eta^{l-1}_{ijk}]$. In fact, $E_D E_{u|D} \left(\frac{\partial^2}{\partial\theta\partial\theta'} l_{ic}(\theta) \right) = B_i$, with $\sup_{i\geq 1} \|B_i\| < \infty$ and D stands for the whole data set.

A.5 True parameter vector $\theta^0 = (\lambda^{0l}, \beta^{0l}, \gamma^{0l})'$ satisfies $\|\theta^0\| \leq M_0$ for some known constant $M_0 (> 0)$.

Assumption A.1 essentially indicates that we have only local dependence in the sample. Assumption A.2 is the compact support for covariates. The smoothness condition on f0 given in assumption A.3 determines the rate of convergence of the spline estimate $\hat{f} = v'(t)\hat{\gamma}$. Both the assumptions A.2 and A.4 are natural and are easy to check. Assumption A.5 is basically a technical condition required to justify consistency.

It is true that, in our model, the covariates x_{ijk} may be time dependent and hence must depend on t_{ijk}. Such dependence can be taken into account through some relationship (either linear or non-linear). For example, we can express covariates as,

$$X_{ijku} = \Psi_u(t_{ijk}) + \varepsilon_{ijku}; i = 1, \ldots, n; j = 1, \ldots, r; k = 1, \ldots, K; u = 1, \ldots, p. \tag{4.1}$$

where $\Phi_u(\cdot)$ are p functions for each of which sth derivative is bounded and ε_{ijku}'s are independent random variables with mean zero and also independent of y_{ijk}'s. In view of the fact that γ is the nuisance parameter vector, for clear representation we modify equation (3.3) as,

$$\hat{\theta} = \hat{\theta}^0 - \left[\Lambda_1^{*-1} \xi_\theta^*(\theta)\right]_{\theta=\hat{\theta}_0} \tag{4.2}$$

where

$$\begin{aligned}
&A_1^* = E\left[\frac{\partial}{\partial\theta'}(X^{*\prime} W Y_0) \mid D\right], \xi_\theta^* = E\left[X^{*\prime} W Y_0 \mid D\right], X^* = (I - H)X, \\
&H = P(P'P)^{-1}P', P = 1_L \otimes v'(t_{ijk}), \\
&Y_0 = (y_{ijkl}, i = 1, \ldots, n; j = 1, \ldots, r; k = 1, \ldots, K; l = 0, \ldots, L-1)', \\
&R_n = (X^{*\prime} X^*), W = \text{ Diag } (\ldots, 1 - \eta'_{ijk} - \eta^{l-1}_{ijk}, \ldots) \text{ and} \\
&X = \begin{pmatrix} 1_L \otimes 1'_L & x'_{111} & v'_{111} \\ 1_L \otimes 1'_L & x'_{112} & v'_{112} \\ \vdots & \vdots & \vdots \\ 1_L \otimes 1'_L & x'_{nrk} & v'_{nrk} \end{pmatrix}
\end{aligned}$$

For the existence of Fisher Information, the following assumptions are further made,

$$\text{A.6 } (i) \lim_{n\to\infty} \frac{k_n}{n}(P'P) = Q, \quad (ii) \lim_{n\to\infty} R_n = R.$$

In assumption, A.6 (i), k_n is the number of knots, Q and Rare positive definite matrices with all eigen values bounded. Assumption A.6 (i) is a very standard property of B-spline basis functions and holds true under general design conditions (He and Shi, 1996). A.6 (ii) is a prerequisite for the existence of asymptotic distribution of the proposed estimator. The asymptotic distribution of $\hat{\beta}_n$ then follows from the following theorem:

Theorem 1. *Under assumptions A.1-A.6, the MLE $\hat{\theta}$ of θ^0 is consistent i.e. $\|\hat{\theta} - \theta^0\| \xrightarrow{P} 0$ as $n \to \infty$. Specifically as $n \to \infty$,*

$$\left(\hat{\beta} - \beta^0\right) \xrightarrow{P} 0, \qquad \sup_{t\in[0,r]} \mid v'(t)\hat{\gamma} - f_0(t) \mid \to 0. \tag{4.3}$$

The sketch of the proof is given in Appendix A2.

5 Simulation Study

In the simulation study we focus on a setting where $L = 4, K = 4, r = 4$ and $n = 100$. We simulate the clustered longitudinal ordinal response from a model with,

$$\text{logit}(\eta'_{ijk}) = \lambda_l + \beta x_i + u_{ij} + \sin(\pi t_{ijk}), \tag{5.1}$$

where the monotone difference intercepts $(\lambda_0, \lambda_1, \lambda_2)$ are assigned the value $(-2.0, -1.5, -1.0)$ and the regression parameter β is chosen to be 0.5. The time dependent covariate t_{ijk} is simulated from Uniform $(-1, 1)$ while the baseline covariate x_i is generated from $N(0, 1)$. The random component $u_i = (u_{i1}, \ldots, u_{ir})'$ is generated from a r-variate normal distribution with mean zero and variance-covariance matrix given by $\sigma_u^2 \left[(\mathbf{1} - \rho)I_r + \rho\mathbf{1}\mathbf{1}'\right]$, where the true values of σ_u^2 and p are taken to be 1.0 and 0.6 respectively. During the estimation process the function $\sin(\pi t_{ijk})$ is approximated by the normalized cubic B spline basis function. The data adaptive scheme outlined in Section 3.2 is applied and the number of internal knots is chosen to be 4. The knot points are taken as the 20th, 40th, 60th and 80th percentile values of $t_{ijk}; i = 1, \ldots, n; j = 1, \ldots, r; k = 1, \ldots, K$. Metropolis Hastings (MH) algorithm is employed for generating observations from the conditional distribution of u_i given the data. For simplicity and time saving purpose, the MH sample size R is chosen to be 500. The number of iterations needed in the Newton Raphson method within the Metropolis algorithm is predetermined to be 30. This resulted in about two-decimal accuracy in the simulation study. The simulation is repeated 100 times. For each parameter θ_i associated with the outcome model the goodness of fit measures namely bias and mean square error (MSE) are computed. Suppose $\hat{\theta}_{u'}$ denote the estimate of θ_i in the $t'th$ simulated data. Then Bias and MSE are

given by,

$$\text{Bias}_i = \frac{1}{100}\sum_{u=1}^{100}(\hat{\theta}_u - \theta_i); MSE_i = \frac{1}{100}\sum_{u=1}^{100}(\hat{\theta}_u - \theta_i)^2. \tag{5.2}$$

The first measure assesses the accuracy of $\hat{\theta}_i$ and the second measure assesses the precision. We also compared the efficiency of the proposed model with the naïve model. For a naïve model the ordinal responses are generated using (5.1), but we fit a model after replacing the nonlinear function of time by t_{ijk} simply. The estimated values of the parameters along with the bias and MSE of the estimates of the parameters are presented in Table 1 for the naïve model as well as for the proposed model which accounts for the longitudinal effect through spline function. The program has been implemented in R 2.14.1.

Table 1: Parameter estimates, simulated biases and mean square error of the parameter estimates for the proposed model and naive model.

Parameters	True	Naive Model			Proposed Model		
	values	Estimates	Bias	MSE	Estimates	Bias	MSE
λ_0	-2.0	-1.8059	0.1933	0.3802	-2.0319	0.0319	0.0435
λ_1	-1.5	-1.419	0.1065	0.3334	-1.507	-0.0906	0.0227
λ_2	-1.0	-1.155	-0.1563	0.3866	-1.003	-0.0032	0.0161
β	0.5	0.5178	0.0178	0.0152	0.4968	-0.0131	0.0097
σ_u^2	1.0	0.9137	-0.0232	0.0128	0.9677	-0.0962	0.0020
ρ	0.6	0.6000	0.0000	.0000	0.6000	0.0000	0.0000

Table 1 shows that the monotone difference estimates and the regression coefficients are biased under the naïve model, whereas the proposed model recovers the estimates well. However the estimates of the parameters associated with the distribution of the random component remains robust under model misspecification. In the naïve model we have 6 parameters while the proposed model involves 13 parameters. The AIC factor under naïve model comes out to be 3642.622 while that under the proposed model is 3552.039. During the estimation process under the proposed model the spline coefficients $\gamma = (\gamma_1, \ldots, \gamma_N)$ are also estimated along with the other parameters of interest. The fitted function is then given by,

$$\hat{f}_0(t) = \sum_{m=1}^{N} \hat{\gamma}_m B_m(t), \tag{5.3}$$

where $\hat{\gamma}_m$ is the estimated value of γ_m and $B_m(t)$ denotes the B-spline basis function. The calculation of basis functions for the cubic B-spline is done using the $\{$splines$\}$ package in R. With four internal knot points and spline of order 3 and intercept =False the $bs(.)$ function in R returns $N = 7$

basis functions. Figure 1 displays the graph of the fitted function given by (5.3) and the true function given by $\sin(\pi t)$against the different values of t_{ijk}. The graph reveals that the cubic B-spline basis function approximates the true function $\sin(\pi t)$ well.

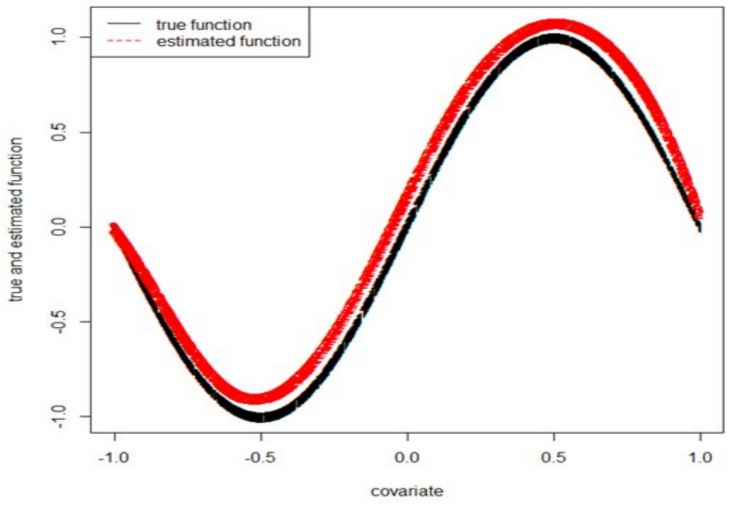

Figure 1: Plot of true function and estimated function against the time- dependent covariate.

For justification of the working of MCMHNR algorithm a simpler set up is chosen. Here we assume that in model $(16), \rho = 0$. This leads to uncorrelated random components and hence the multidimensional integration over $u_i = (u_{i1}, u_{i2}, u_{i3}, u_{i4})'$ is reduced to one dimensional integrals. The score equations now involve integrals over the random component u_i, which are evaluated using Gauss Hermite quadrature. Alternatively we apply MCMHNR algorithm as outlined in Section 3 under this simple set up. The likelihood estimates of the parameters are computed for each case. The AIC for the exact approach comes out as 2641.332, while that on application of EM algorithm is 2599.236. This shows that the MCMHNR method approximates the exact likelihood approach well.

6 Data Analysis

In this section we motivate the proposed model through an analysis of orthodontic data. Oral hygiene is of severe concern as a significant proportion of world population is highly susceptible to some destructive periodontal diseases. The data are the result of a study of 220 individuals consisting of staff members and students of medical schools in and around the city of Kolkata. These individuals have been selected at random irrespective of age, gender and oral hygiene status. A detailed history of each subject was recorded a week prior to the beginning of the study to collect information

like age, gender, occupation, food habits and smoking habits. Plaque scoring was done according to Tureskey et al. (1970). The teeth selected for scoring of plaque were the maxillary right first permanent molar, maxillary left permanent central incisor, maxillary left first premolar, mandibular left first permanent molar, mandibular right central incisor and mandibular right first premolar which we shall denote as teeth 1-6. Ordinal score of 0-2 was assigned as: 0 (No plaque), 1 (A thin band of plaque up to 1 mm at the cervical margin of the crown of the tooth.), 2 (A band of plaque wider than 1 mm of the crown of the tooth).

The categories 'moderate reduction' and 'vast reduction' were assigned the ordinal scores 1 and 2 respectively while the categories 'no reduction' and 'slight reduction' were combined and given the ordinal score 0.The plaque scoring on individual teeth was measured on four occasions separated at an interval of 1 month. Figure 2 shows the average response (plaque score) over time for each of the six teeth. The graph reveals a non-linear pattern in plaque deposit over time. The main focus of this orthodontic study is to see whether plaque reduction is truly effective with the use of a solution (kept in mouth for 1 minute followed by a thorough rinse with water to remove any excess of disclosing solution) and if so, to what extent such progression (i.e. plaque reduction) depends on the covariates taken. We consider the following model:

$$\eta^l_{ijk} = \lambda_l + \beta_A x_{Ai} + \beta_G x_{Gi} + \beta_F x_{Fi} + \beta_S x_{Si} + u_{ij} + f_0(t_K). \tag{6.1}$$

In equation (6.1) above, the baseline covariates $x_{Ai},\ x_{Gi},\ x_{Fi},\ x_{Si}\ \ (i = 1, \ldots, 220)$ correspond to age, gender, food habit and smoking habit respectively. The binary covariates x_{Gi}, x_{Fi} and x_{Si} takes the value 1 if the person is a male , non-vegetarian and a smoker. The non-linear behavior of the response over time is captured by the smooth unknown function $f_0(t_k)(k = 1, \ldots, 4)$, where $t_k = k$. In the analysis, the unknown function is approximated by a smoothing spline of order 1 with 4 internal knot points. Table 2 provides the estimated values of the parameters along with their standard errors for both the naive model and the proposed model. The naive model replaces the non-linear function by t_k. The results reveal that the smokers will have on an average less value of the response i.e. plaque reduction. Moreover food habit is not a significant factor in determining the effect of the solution (treatment) on plaque reduction.In this study 'age' does not play a significant role. The reason for this may be that the subjects considered belonged to almost the same age group. Finally it can be concluded from the results that the particular treatment applied on plaque reduction had better effect on males. The fitted function $\hat{f}_0(t) = \sum_{m=1}^{N} \hat{\gamma}_m B_m(t)$ is computed for $N = 3$. Here $\hat{\gamma}_m$ denotes the estimated value of the spline coefficients corresponding to the basis spline function $B_m(t)$ for different time points (t). The function shows a non-linear decreasing trend over time. Thus it can be inferred that in general the application of the solution helps in reducing plaque deposit over time.

7 Conclusion

In many longitudinal set up where responses are ordinal in nature, one faces the stiff challenge in expressing the dependence of such responses over time. In our present orthodontic study, it is evident from Figure 2 that average response (plaque reduction) varies nonlinearly over time. The variation

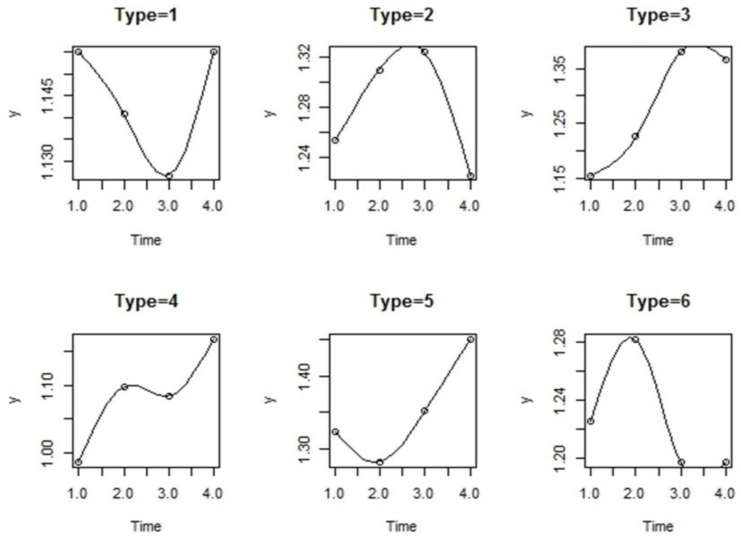

Figure 2: The average response (plaque score) over time for each of the six teeth.

also changes over the six teeth. To account for such unknown variability, we have proposed a GPOLM that can be viewed as a compromise between GLMM and a fully nonparametric model. We have approximated the non-parametric function in the GPOLM by a regression spline. A MCMHNR method has been proposed to estimate the model parameters. Simulation study indicates that the model which ignores the non-linear effect of time produces biased estimates of the intercepts and the regression coefficients. Result from the orthodontic study reveals that smoking has a negative effect in plaque reduction. However in general the application of the solution helps in reducing plaque deposit over time.

Acknowledgment

The authors are thankful to the reviewer for careful suggestions that helped to improve the clarity of the manuscript.

Appendix A1

First order derivatives:

$$\frac{\partial l_{c_1}(\theta)}{\partial \theta} = \sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{K}\sum_{l=1}^{L} \frac{I\left(y_{ijk}=l\right)}{\left(\eta_{ijk}^{l}-\eta_{ijk}^{l-1}\right)}\left(\frac{\partial \eta_{ijk}^{l}}{\partial \theta}-\frac{\partial \eta_{ijk}^{l-1}}{\partial \theta}\right), \qquad (A1.1)$$

Table 2: Estimated values of the covariate effects along with their standard errors.

Parameters	Naive Model		Proposed Model	
	Estimates	Standard Error	Estimates	Standard Error
λ_0	-3.091	0.2211	-3.0624	0.2186
λ_1	0.4670	0.1237	0.6944	0.1172
β_{FOOD}	0.0956	0.2082	0.0095	0.1602
β_{AGE}	-0.0030	0.1102	-0.0004	0.0080
β_{GENDER}	0.3095	0.2113	0.3744	0.1865
β_{SMOKE}	-0.1651	0.1619	-0.3023	0.1510
σ_u^2	0.9077	0.1153	0.9513	0.0629
ρ	0.6002	0.0014	0.6000	0.0014

where $I(x)$ is an indicator function, $\eta'_{ijk} =$ logit $(\lambda_1 + x'_{ijk}\beta + z'_{ijk}u_{ij} + v'(t_{ijk})\gamma)$, $\theta = (\lambda_0, \ldots, \lambda_{L-1}, \beta', \gamma')'$, and

$$\left.\begin{array}{rcl} \frac{\partial \eta'_{ijk}}{\partial \lambda_l} & = & \eta'_{ijk}(1-\eta'_{ijk}), \;\; l = 0, \ldots, L-1, \\ \frac{\partial \eta'_{ijk}}{\partial \beta'} & = & \eta'_{ijk}(1-\eta'_{ijk})x'_{ijk} \\ \frac{\partial \eta'_{ijk}}{\partial \gamma'} & = & \eta'_{ijk}(1-\eta'_{ijk})v'(t_{ijk}) \end{array}\right\} \qquad (A1.2)$$

Substituting (A1.2) in (A1.1) we get,

$$\frac{\partial l_{c_1}(\theta)}{\partial \theta} = \sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{K}\sum_{l=1}^{L} I\left(y_{ijk} = l\right)\left(1 - \eta'_{ijk} - \eta_{ijk}^{l-1}\right)\tilde{X}_{ijk}, \qquad (A1.3)$$

where $\tilde{X}_{ijk} = (1' x'_{ijk} v'(t_{ijk}))'$.

Second order derivatives:

$$\frac{\partial^2(\theta)}{\partial\lambda_l^2} = \sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{K}\sum_{l=1}^{L} I\left(y_{ijk}=l\right) b'_{ijk}; l=0,\ldots,L-1$$

$$\frac{\partial^2 l_{c_1}(\theta)}{\partial\beta\partial\beta'} = \sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{K}\sum_{l=1}^{L} I\left(y_{ijk}=l\right) b'_{ijk} x_{ijk} x'_{ijk}$$

$$\frac{\partial^2 l_{c_1}(\theta)}{\partial\gamma\partial\gamma'} = \sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{K}\sum_{l=1}^{L} I\left(y_{ijk}=l\right) b'_{ijk} v(t_{ijk}) v'(t_{ijk})$$

$$\frac{\partial^2 l_{c_1}(\theta)}{\partial\beta^T\partial\lambda_l} = \sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{K}\sum_{l=1}^{L} I\left(y_{ijk}=l\right) b'_{ijk} x'_{ijk}$$

$$\frac{\partial^2 l_{c_1}(\theta)}{\partial\gamma'\partial\lambda_l} = \sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{K}\sum_{l=1}^{L} I\left(y_{ijk}=l\right) b'_{ijk} v'(t_{ijk})$$

$$\frac{\partial^2 l_{c_1}(\theta)}{\partial\beta'\partial\gamma} = \sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{K}\sum_{l=1}^{L} I\left(y_{ijk}=l\right) b'_{ijk} x_{ijk} v'(t_{ijk})$$

where $b'_{ijk} = -\eta'_{ijk}(1-\eta'_{ijk}) - \eta_{ijk}^{l-1}(1-\eta_{ijk}^{l-1})$

Appendix A2

Proof of Theorem 1 : We give an outline of the proof as it is essentially based on the result of Stone (1985). Equation (4.3) can be proved following Lemma 8 and 9 in Stone (1985). In fact, it can be shown that if the number of knots $k_n \cong O\left(n^{\frac{1}{(2m+1)}}\right)$ then for $m \geq 2$,

$$\frac{1}{nrK}\sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{K}\left(v'(t_{ijk})\hat{\gamma} - f_0(t_{ijk})\right)^2 = O_P\left(n^{\frac{-2m}{(2m+1)}}\right) \qquad (A2.1)$$

Expression (A2.1), in view of Stone (1985) can be expressed as,

$$\int\left\{\hat{f}(t) - f_0(t)\right\}^2 dt = O_P\left(n^{\frac{-2m}{(2m+1)}}\right) \qquad (A2.2)$$

The proof of equations (4.3) are rather straightforward application of Zeng and Cai (2005). Under assumptions A.1–A.6 a solution to equation (3.2) exists and with probability unity, $\hat{\theta} \to \theta^0$.

References

[1] Akaike H (1973). Information theory and an extension of the maximum likelihood principle. *In second International Symposium on Information Theory (BN Petrov and F Csaki, eds.)*. Akademiai Kiado, Budapest.

[2] Breslow NE and Clayton DG (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**(421), 9–25.

[3] Breslow NE and Lin X (1995). Bias correction generalized linear models with single component dispersion. *Biometrika* **82**(1), 81–92.

[4] Celeux G, Chauveau D and Diebolt J (1995). On Stochastic versions of the EM algorithm. *INRIA- Rapport de recherch?* **2514**.

[5] Chib S and Greenberg E (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**(4), 327–335.

[6] Davier MV and Sinharay S (2010). Stochastic Approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics*, **35** (2), 174–193.

[7] Delyon B, Lavielle M and Moulines E (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, **27**(1), 94–128.

[8] de Boor C (2001). *A practical guide to splines*. Spring-Verlag.

[9] Greene W. H. (2004). *Econometric Analysis*. Prentice Hall.

[10] Hardle W, Mammen E and Muller M (1998). Testing parametric versus semiparametric modelling in generalized linear models. *Journal of the American Statistical Association*, **93**(444), 1461–1474.

[11] Hardle W, Liang H and Gao J (2000). *Partially linear models*. *Physica-Verlag*.

[12] He X and Shi PD (1996). Bivariate tensor product b-spline in a partly linear model. *Journal of Multivariate Analysis*, **58**(2), 162–181.

[13] He X, Fung WK and Zhu Z (2005). Robust estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association*, **100**(472), 1176–1184.

[14] Li CS (2011). A lack-of-fit test for parametric zero-inflated Poisson models. *Journal of Statistical Computation and Simulation* **81**(9), 1081–1098.

[15] Lin X and Breslow NE (1996). Bias correction generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, **91**(435), 1007–1016.

[16] Lin X and Carroll RJ (2001a). Semiparametric regression for clustered data. *Biometrics*, **88**(4), 1179–1185.

[17] Lin X and Carroll RJ (2001b). Semiparametric regression for clustered data using generalized estimating equations. *Journal of American Statistical Association*, **99**(466), 1045–1056.

[18] McCullagh P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society B*, **42**(2), 109–142.

[19] McCulloch CE (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89**(428), 330–335.

[20] McCulloch CE (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**(437), 162–170.

[21] Meza C, Jaffrezic F and Foulley JL (2009). Estimation in the probit normal model for binary outcomes using the SAEM algorithm. *Computational Statistics and Data Analysis*, **53**(4), 1350–1360.

[22] Muller M (2001). Estimation and testing in generalized partial linear models - a comparative study. *Statistics and Computing*, **11**(4), 299–309.

[23] Natarajan R, McCulloch CE and Kiefer NM (2000). A Monte Carlo EM method for estimating multinomial probit models. *Computational Statistics and Data Analysis*, **34**(1), 33–50.

[24] Pinheiro JC and Bates DM (1995). Approximations to the log-likelihood function in nonlinear mixed-effects models. *Journal of Computational Graphical Statistics*, **4**(1), 12–35.

[25] Severini TA and Staniswalis JG (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, **89**(428), 501–511.

[26] Sinha SK (2004). Robust analysis of generalized linear mixed models. *Journal of the American Statistical Association*, **99**(466), 451–460.

[27] Smith AFM and Roberts GO (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, **55**(1), 3–24.

[28] Stone C(1985). Additive regression and other nonparametric models. *Annals of Statistics*, **13**(2), 689–705.

[29] Turskey S, Gilmore ND and Glickman IR (1970). Reduced plaque formation by the chloromethyl analogue of Vit-C. *Journal of Periodontology*, **41**(1), 41–43.

[30] Wang N, Carroll RJ and Lin X (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, **100**(469), 147–157.

[31] Wooldridge J (2003). *Econometric analysis of cross section and panel data*. MIT press.

[32] Zeger SL and Karim MR (1991). Generalized linear model with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**(413), 79–86.

[33] Zeng D and Cai J (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Annals of Statistics*, **33**(5), 2132–2163.

[34] Zhou X and Liu X (2008). The Monte Carlo EM method for estimating multinomial probit latent variable models. *Computational Statistics*, **23**(2), 277–289.

Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 1, pp. 35-46

A BIVARIATE VERSION OF THE HYPER-POISSON DISTRIBUTION AND SOME OF ITS PROPERTIES

C. SATHEESH KUMAR
Department of Statistics, University of Kerala, Trivandrum, India.
Email: drcsatheeshkumar@gmail.com

B. UNNIKRSHNAN NAIR
University of Kerala, Trivandrum, India.
Email: ukswathy@gmail.com

SUMMARY

A bivariate version of the hyper-Poisson distribution is introduced here through its probability generating function (*pgf*). we study some of its important aspects by deriving its probability mass function, factorial moments, marginal and conditional distributions and obtain certain recurrence relations for its probabilities, raw moments and factorial moments. Further, the method of maximum likelihood is discussed and the procedures are illustrated using a real life data set.

Keywords and phrases: Confluent hypergeometric function, Displaced Poisson distribution, Factorial moment generating function, Hermite distribution, Poisson distribution.

AMS Classification: Primary 60E05, 60E10; Secondary 33C20.

1 Introduction

Bivariate discrete distributions have received a great deal of attention in the literature. For details see Kumar (2008), Kocherlakota and Kocherlakota (1992) and references therein. Bardwell and Crow (1964) studied the hyper-Poisson distribution (HP distribution), which they defined as follows. A random variable X is said to follow an HP distribution if it has the following probability mass function (*pmf*) , for $x = 0, 1, \ldots$

$$g(x) = P(X = x) = \frac{\theta^x \, \Gamma(\lambda)}{\phi(1; \lambda; \theta) \, \Gamma(\lambda + x)}, \tag{1.1}$$

in which λ, θ are positive real numbers and $\phi(1; \lambda; \theta)$ is the confluent hypergeometric series (for details see Mathai and Saxena, 1973 or Slater, 1960). The probability generating function (*pgf*) of the HP distribution with *pmf* (1.1) is the following

$$G(t) = \frac{\phi(1; \lambda; \theta t)}{\phi(1; \lambda; \theta)}, \tag{1.2}$$

which reduces to Poisson distribution when $\lambda = 1$ and when λ is a positive integer, the distribution is known as the displaced Poisson distribution studied by Staff (1964). Bardwell and Crow (1964) termed the distribution as sub-Poisson when $\lambda < 1$ and super-Poisson when $\lambda > 1$. Bardwell and Crow (1964) and Crow and Bardwell (1965) considered various methods of estimation of the parameters of the distribution. Some queuing theory with hyper-Poisson arrivals has been developed by Nisida (1962) and certain results on moments of hyper-Poisson distribution has been studied in Ahmad (1979). Roohi and Ahmad (2003a) discussed the estimation of the parameters of the hyper-Poisson distribution using negative moments. Roohi and Ahmad (2003b) obtained certain recurrence relations for negative moments and ascending factorial moments of the HP distribution. Kemp (2002) developed q-analogue of the HP distribution and Ahmad (2007) introduced and studied Conway-Maxwell hyper-Poisson distribution. Kumar and Nair (2011, 2012a, 2012b) introduced modified versions of the HP distribution and discussed some of their applications.

Ahmad (1981) introduced a bivariate version of the HP distribution through the following *pgf*

$$Q(t_1, t_2) = (\phi_1 \phi_2)^{-1} \exp[\theta(t_1 - 1)(t_2 - 1)]\phi_1[1; \lambda_1; \theta_1 t_1]\phi_2[1; \lambda_2; \theta_2 t_2], \tag{1.3}$$

in which $\phi_i = \phi(1; \lambda_i; \theta_i)$. For $r \geq 0$, $s \geq 0$, the *pmf* $q(r, s) = P(Z_1 = r, Z_2 = s)$ of $Z = (Z_1, Z_2)$ with *pgf* (1.3) is the following

$$q(r,s) = \frac{e^{\theta}\Gamma(\lambda_1)\Gamma(\lambda_2)}{\phi_1\,\phi_2} \sum_{i=0}^{\min(r,s)} \sum_{j=0}^{r-i} \sum_{k=0}^{s-i} \frac{(-1)^{j+k}\theta_1^{r-i-j}\theta_2^{s-i}\theta^{i+j+k}}{\Gamma(\lambda_1 + r - i - j)\Gamma(\lambda_2 + s - i - k)i!j!k!}, \tag{1.4}$$

where $\lambda_1 > 0$, $\lambda_2 > 0$ and $0 < \theta \leq min(\theta_1/\lambda_1, \theta_2/\lambda_2)$.

Through the present paper we introduce another bivariate version of the HP distribution, which we named as 'the bivariate hyper-Poisson distribution ($BHPD$)'and obtain its important properties. In section 2, it is shown that the $BHPD$ possess a random sum structure. Further we obtain its conditional probability distribution, probability mass function and factorial moments in section 2. In section 3, we develop certain recursion formulae for probabilities, raw moments and factorial moments of the $BHPD$ and in section 4 we discuss the estimation of the parameters of the $BHPD$ by the method of maximum likelihood and the distribution has been fitted to a well-known data set and it is observed that the $BHPD$ gives better fit than the bivariate Poisson distribution and the bivariate hyper-Poisson distribution of Ahmad (1981).

Note that the bivariate version of HP distribution introduced in this paper is relatively simple in terms of its *pmf* and *pgf* compared to the bivariate version due to Ahmad(1981), and further this bivariate form possess a bivariate random sum structure as given in section 2. The random sum structure arises in several areas of research such as ecology, biology, genetics, physics, operation research etc. For details, see Johnson et al. (2005).

2 The BHP distribution

Consider a non-negative integer valued random variable X following HP distribution with *pgf* (1.2), in which $\theta = \theta_1 + \theta_2 + \theta_3$, $\theta_1 > 0$, $\theta_2 > 0$ and $\theta_3 \geq 0$. Define $\alpha_j = \theta_j/\theta$, for $j = 1, 2, 3$ and let

$\{Y_n = (Y_{1n}, Y_{2n}),\ n = 1, 2, \ldots\}$ be a sequence of independent and identically distributed bivariate Bernoulli random vectors, each with *pgf*

$$P(t_1, t_2) = \alpha_1 t_1 + \alpha_2 t_2 + \alpha_3 t_1 t_2.$$

Assume that $X, Y_1, Y_2, \ldots$ are independent. Let $T_0 = (T_{10}, T_{20}) = (0, 0)$ and define

$$T_X = (T_{1X}, T_{2X}) = \left(\sum_{x=1}^{X} Y_{1x}, \sum_{x=1}^{X} Y_{2x} \right).$$

Then the *pgf* of T_X is the following, in which $\Lambda = \phi^{-1}(1; \lambda; \theta_1 + \theta_2 + \theta_3)$.

$$H(t_1, t_2) = G\{P(t_1, t_2)\} = \Lambda\phi(1; \lambda; \theta_1 t_1 + \theta_2 t_2 + \theta_3 t_1 t_2) \tag{2.1}$$

We call a distribution with *pgf* as given in (2.1) as 'the bivariate hyper-Poisson distribution'or in short, 'the $BHPD$'. Clearly the $BHPD$ with $\lambda = 1$ is the bivariate Poisson distribution discussed in Kocherlakotta and Kocherlakotta (1992, pp 90) and when λ is a positive integer the $BHPD$ with *pgf* (2.1) reduces to the *pgf* of a bivariate version of the displaced Poisson distribution.

Let (X_1, X_2) be a random variable having the $BHPD$ with *pgf* (2.1). Then the marginal *pgf* of X_1 and X_2 are respectively

$$H_{X_1}(t) = H(t, 1) = \Lambda\phi[1; \lambda; (\theta_1 + \theta_3)t + \theta_2] \text{ and}$$
$$H_{X_2}(t) = H(1, t) = \Lambda\phi[1; \lambda; (\theta_2 + \theta_3)t + \theta_1].$$

The *pgf* of $X_1 + X_2$ is

$$H_{X_1+X_2}(t) = H(t, t) = \Lambda\phi[1; \lambda; (\theta_1 + \theta_2)t + \theta_3 t^2],$$

which is the *pgf* of a modified version of the HP distribution studied in Kumar and Nair (2011).

Let x be a non-negative integer such that $P(X_2 = x) > 0$. On differentiating (2.1) with respect to $t_2 x$ times and putting $t_1 = t$ and $t_2 = 0$, we get

$$H^{(0,x)}(t, 0) = (\theta_2 + \theta_3 t)^x \left(\prod_{j=0}^{x-1} D_j \right) \Lambda\delta_x(\theta_1 t) \tag{2.2}$$

where $D_j = (1 + j)/(\lambda + j)$ and $\delta_j(t) = \phi(1 + j; \lambda + j; t)$ for $j = 0, 1, 2, \ldots$.

Now applying the formula for the *pgf* of the conditional distribution in terms of partial derivatives of the joint *pgf*, developed by Subrahmaniam (1966), we obtain the conditional *pgf* of X_1 given $X_2 = x$ as

$$H_{X_1|X_2=x}(t) = \left(\frac{\theta_2 + \theta_3 t}{\theta_2 + \theta_3} \right)^x \frac{\phi(1 + x; \lambda + x; \theta_1 t)}{\phi(1 + x; \lambda + x; \theta_1)} = H_1(t) H_2(t), \tag{2.3}$$

where $H_1(t)$ is the *pgf* of a binomial random variable with parameters x and $p = \theta_3(\theta_2 + \theta_3)^{-1}$ and $H_2(t)$ is the *pgf* of a random variable following the HPD with parameters $1 + x$, $\lambda + x$ and

θ_1. Note that, when $\theta_3 = 0$ and/or when $x = 0$, $H_1(t)$ reduces to the *pgf* of a random variable degenerate at zero. Thus the conditional distribution X_1 given $X_2 = x$ given in (2.4) can be viewed as the distribution of the sum of independent random variables V_1 with *pgf* $H_1(t)$ and V_2 with *pgf* $H_2(t)$. Consequently from (2.4) we obtain the following

$$E(X_1 \,|X_2 = x\,) = \frac{x\theta_3}{(\theta_2+\theta_3)} + \frac{\theta_1 D_x \delta_{x+1}(\theta_1)}{\delta_x(\theta_1)} \tag{2.4}$$

$$\begin{aligned}\text{Var}(X_1 \,|X_2 = x) \;=\; & \frac{x\theta_2\theta_3}{(\theta_2+\theta_3)^2} + \frac{\theta_1 D_x}{\delta_x^2(\theta_1)}\{D_{x+1}\delta_x(\theta_1)\delta_{x+2}(\theta_1)\theta_1 \\ & + \delta_x(\theta_1)\delta_{x+1}(\theta_1) - D_x[\delta_{x+1}(\theta_1)]^2\theta_1\}.\end{aligned} \tag{2.5}$$

In a similar approach, for a non-negative integer x with $P(X_1 = x) > 0$, we can obtain the conditional *pgf* of X_2 given $X_1 = x$ by interchanging θ_1 and θ_2 in (2.3). Therefore it is evident that comments similar to those in case of the conditional distribution of X_1 given $X_2 = x$ are valid regarding conditional distribution of X_2 given $X_1 = x$ and explicit expressions for $E(X_2 \,|X_1 = x\,)$ and $Var(X_2 \,|X_1 = x)$ can be obtained by interchanging θ_1 and θ_2 in the right hand side expressions of (2.5) and (2.6) respectively.

In order to obtain the probability mass function *pmf* of the $BHPD$, we need the following partial derivatives of $H(t_1, t_2)$, in which r is a non-negative integer.

$$H^{(r,0)}(t_1,t_2) = (\prod_{i=0}^{r-1} D_i)(\theta_1+\theta_3 t_2)^r \Lambda \Delta_r(t_1,t_2), \tag{2.6}$$

where

$$\Delta_j(t_1,t_2) = \phi(1+j;\lambda+j;\theta_1 t_1+\theta_2 t_2+\theta_3 t_1 t_2), j = 0,1,2,\ldots$$

The following derivatives are needed in the sequel, in which $0 \le i \le r$ and $j \ge 1$.

$$\frac{\partial^i(\theta_1+\theta_3 t_2)^r}{\partial t_2^i} = \frac{r!\theta_3^i}{(r-i)!}(\theta_1+\theta_3 t_2)^{r-i} \tag{2.7}$$

$$\frac{\partial^j \Delta_r(t_1,t_2)}{\partial t_2^j} = \prod_{i=r}^{r+j-1} D_i(\theta_2+\theta_3 t_1)^j \Delta_{r+j}(t_1,t_2). \tag{2.8}$$

Differentiating both sides of (2.7) s-times with respect to t_2 and applying (2.8) and (2.9), we get the following

$$\begin{aligned} H^{(r,s)}(t_1,t_2) &= (\prod_{i=0}^{r-1} D_i)\Lambda \sum_{m=0}^{s} \binom{s}{m} \frac{\partial^m(\theta_1+\theta_3 t_2)^r}{\partial t_2^m} \frac{\partial^{s-m}\Delta_r(t_1,t_2)}{\partial t_2^{s-m}} \\ &= (\prod_{i=0}^{r-1} D_i)\Lambda \sum_{m=0}^{\min(r,s)} \binom{s}{m} \frac{r!}{(r-m)!}\theta_3^m(\theta_1+\theta_3 t_2)^{r-m} \\ &\quad \times \prod_{i=r}^{r+s-m-1} D_i(\theta_2+\theta_3 t_1)^{s-m}\Delta_{r+s-m}(t_1,t_2) \end{aligned} \tag{2.9}$$

Now, by putting $(t_1, t_2) = (0, 0)$ in (2.10) and by dividing $r!s!$, we get the *pmf* of the $BHPD$ as

$$h(r,s) = \Lambda\theta_1^r\theta_2^s \sum_{m=0}^{\min(r,s)} \frac{D^*}{m!(r-m)!(s-m)!}(\frac{\theta_3}{\theta_1\theta_2})^m, \tag{2.10}$$

where

$$D^* = \prod_{j=0}^{r+s-m-1} D_j \text{ and } \prod_{j=0}^{k} D_j = 1, \text{for any } k < 0.$$

By putting $(t_1, t_2) = (1, 1)$ in (2.10) we get the $(r, s)^{th}$ factorial moment $\mu_{[r,s]}$ of the $BHPD$ as

$$\mu_{[r,s]} = \Lambda r!s!(\theta_1+\theta_3)^r(\theta_2+\theta_3)^s \sum_{m=0}^{\min(r,s)} \frac{D^*\xi_{r+s-m}}{m!(r-m)!(s-m)!}\beta^m \tag{2.11}$$

where $\xi_j = \phi(1+j; \lambda+j; \theta_1+\theta_2+\theta_3)$, for $j = 0, 1, \ldots$ and $\beta = \theta_3(\theta_1+\theta_3)^{-1}(\theta_2+\theta_3)^{-1}$.

From (2.12) we have the following, in which $\psi_j = \Lambda\xi_j$, for $j = 1, 2, \ldots$

$$E(X_1) = \mu_{[1,0]} = D_0\psi_1(\theta_1+\theta_3) \tag{2.12}$$

$$E(X_2) = \mu_{[0,1]} = D_0\psi_1(\theta_2+\theta_3) \tag{2.13}$$

$$\mathrm{Cov}(X_1, X_2) = D_0(D_1\psi_2 - D_0\psi_1^2)(\theta_1+\theta_3)(\theta_2+\theta_3) + D_0\psi_1\theta_3 \tag{2.14}$$

where D_0 and D_1 are as given in (2.2).

3 Recurrence relations

Let (X_1, X_2) be a random vector following the $BHPD$ with *pgf* (2.1). For j=0, 1, 2, ..., define $\lambda^* + j = (1+j, \lambda+j)$ and $\lambda^{(j)} = (1+j)(\lambda+j)^{-1}$ Now, the *pmf* $h(r, s)$ of the $BHPD$ given in (2.11) we denote by $h(r, s; \lambda^*)$. Then we have the following result in the light of relations:

$$H(t_1, t_2) = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} h(r,s;\lambda^*)t_1^r t_2^s = \Lambda\phi(1;\lambda;\theta_1t_1+\theta_2t_2+\theta_3t_1t_2) \tag{3.1}$$

and

$$\xi_1 \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} h(r,s;\lambda^*+1)t_1^r t_2^s = \phi(2;\lambda+1;\theta_1t_1+\theta_2t_2+\theta_3t_1t_2), \tag{3.2}$$

in which ξ_1 is as given in (2.12).

Result 3.1. The probability mass function $h(r, s; \lambda^*)$ of the $BHPD$ satisfies the following recur-

rence relations.

$$h(r+1,0;\lambda^*) = \frac{D_0\psi_1\theta_1}{(r+1)}h(r,0;\lambda^*+1), r \geq 0 \quad (3.3)$$

$$h(r+1,s;\lambda^*) = \frac{D_0\psi_1}{(r+1)}[\theta_1 h(r,s;\lambda^*+1) + \theta_3 h(r,s-1;\lambda^*+1], r \geq 0, s \geq 1 \quad (3.4)$$

$$h(0,s+1;\lambda^*) = \frac{D_0\psi_1\theta_2}{(s+1)}h(0,s;\lambda^*+1), s \geq 0 \quad (3.5)$$

$$h(r,s+1;\lambda^*) = \frac{D_0\psi_1}{(s+1)}[\theta_2 h(r,s;\lambda^*+1) + \theta_3 h(r-1,s;\lambda^*+1], r \geq 1, s \geq 0 \quad (3.6)$$

Proof. Relation (2.7) with r = 1 gives

$$H^{(1,0)}(t_1,t_2) = D_0(\theta_1+\theta_3 t_2)\Lambda\Delta_1(t_1,t_2) \quad (3.7)$$

On differentiating both sides of (3.1) with respect to t_1, we have

$$H^{(1,0)}(t_1,t_2) = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}(r+1)h(r+1,s;\lambda^*)t_1^r t_2^s \quad (3.8)$$

By using (3.2) and (3.8) in (3.7) we get the following, in which ψ_1 is as given in (2.13).

$$\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}(r+1)h(r+1,s;\lambda^*)t_1^r t_2^s = D_0\psi_1[\theta_1\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}h(r,s;\lambda^*+1) t_1^r t_2^s + \theta_3\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}h(r,s;\lambda^*+1)t_1^r t_2^{s+1}] \quad (3.9)$$

On equating the coefficient of $t_1^r t_2^0$ on both sides of (3.9) we get the relation (3.3) and on equating the coefficient of $t_1^r t_2^s$ on both sides of (3.9) we get the relation (3.4). We omit the proof of relations (3.5) and (3.6) as it is similar to that of relations (3.3) and (3.4). □

Result 3.2. For $r, s \geq 0$, simple recurrence relations for factorial moments $\mu_{[r,s]}(\lambda^*)$ of order (r, s) of the $BHPD$ are the following.

$$\mu_{[r+1,s]}(\lambda^*) = D_0\psi_1(\theta_1+\theta_3)\mu_{[r,s]}(\lambda^*+1) + D_0\psi_1\theta_3 s\mu_{[r,s-1]}(\lambda^*+1) \quad (3.10)$$

$$\mu_{[r,s+1]}(\lambda^*) = D_0\psi_1(\theta_2+\theta_3)\mu_{[r,s]}(\lambda^*+1) + D_0\psi_1\theta_3 r\mu_{[r-1,s]}(\lambda^*+1), \quad (3.11)$$

in which $\mu_{[0,0]}(\lambda^*) = 1$.

Proof. Let (X_1, X_2) be a random vector having the $BHPD$ with *pgf* $H(t_1,t_2)$ as given in (2.1). Then the factorial moment generating function $F(t_1,t_2)$ of the $BHPD$ is

$$\begin{aligned} F(t_1,t_2) &= H(1+t_1,1+t_2) \\ &= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\mu_{[r,s]}(\lambda^*)\frac{t_1^r t_2^s}{r!s!} \\ &= \Lambda\phi[1;\lambda;\theta_1+\theta_2+\theta_3+(\theta_1+\theta_3)t_1+(\theta_2+\theta_3)t_2+\theta_3 t_1 t_2] \end{aligned} \quad (3.12)$$

Differentiate (3.12) with respect to t_1 to get

$$\begin{aligned}\frac{\partial F(t_1,t_2)}{\partial t_1} &= [(\theta_1+\theta_3)+\theta_3 t_2)]D_0\Lambda \\ &\times \quad \phi[2;\lambda+1;\theta_1+\theta_2+\theta_3+(\theta_1+\theta_3)t_1+(\theta_2+\theta_3)t_2+\theta_3 t_1 t_2] \quad (3.13)\end{aligned}$$

Based on the similar argument as in the proof of Result 3.1., by using (3.12) with λ^* replaced by λ^*+1, one can obtain the following from (3.13).

$$\begin{aligned}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\mu_{[r+1,s]}(\lambda^*)\frac{t_1^r t_2^s}{r!s!} &= D_0\psi_1\{(\theta_1+\theta_3)\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\mu_{[r,s]}(\lambda^*+1)\frac{t_1^r t_2^s}{r!s!} \\ &+ \theta_3\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\mu_{[r,s]}(\lambda^*+1)\frac{t_1^r t_2^{s+1}}{r!s!}\} \quad (3.14)\end{aligned}$$

Now on equating the coefficients of $(r!s!)^{-1}t_1^r t_2^s$ on both sides of (3.14) we obtain the relation (3.10). A similar procedure implies (3.11). □

Result 3.3. Two recurrence relations for the (r, s)th raw moments $\mu_{r,s}(\lambda^*)$ of the $BHPD$ are:

$$\mu_{r+1,s}(\lambda^*) = D_0\psi_1\theta_1\sum_{j=0}^{r}\binom{r}{j}\mu_{r-j,s}(\lambda^*+1) + D_0\psi_1\theta_3\sum_{j=0}^{r}\sum_{k=0}^{s}\binom{r}{j}\binom{s}{k}\mu_{r-j,s-k}(\lambda^*+1) \quad (3.15)$$

and

$$\mu_{r,s+1}(\lambda^*) = D_0\psi_1\theta_2\sum_{k=0}^{s}\binom{s}{k}\mu_{r,s-k}(\lambda^*+1) + D_0\psi_1\theta_3\sum_{j=0}^{r}\sum_{k=0}^{s}\binom{r}{j}\binom{s}{k}\mu_{r-j,s-k}(\lambda^*+1) \quad (3.16)$$

Proof. The characteristic function $A(t_1,t_2)$ of the $BHPD$ with *pgf* (2.1) is the following. For (t_1,t_2) in R^2,

$$A(t_1,t_2) = H(e^{it_1},e^{it_2}) = \Lambda\phi[1;\lambda^*;\lambda(t_1,t_2;\theta)] = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\mu_{r,s}(\lambda^*)\frac{(it_1)^r(it_2)^s}{r!s!}, \quad (3.17)$$

where $\lambda(t_1,t_2;\theta) = \theta_1 e^{it_1}+\theta_2 e^{it_2}+\theta_3 e^{i(t_1+t_2)}$, $\theta=(\theta_1,\theta_2,\theta_3)$ and $i=\sqrt{-1}$. On differentiating (3.17) with respect to t_1, we obtain

$$D_0\Lambda\phi[2;\lambda^*+1;\lambda(t_1,t_2;\theta)]\{i(\theta_1+\theta_3 e^{it_2})e^{it_1}\} = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}i\mu_{r,s}(\lambda^*)\frac{(it_1)^{r-1}(it_2)^s}{(r-1)!s!}.$$

By using (3.17) with λ^* replaced by λ^*+1; and on expanding the exponential functions, we obtain the following, in the light of some standard properties of double sum

$$\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{\mu_{r+1,s}(\lambda^*)(it_1)^r(it_2)^s}{r!s!}$$
$$= D_0\psi_1\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{(it_1)^r(it_2)^s}{r!s!}\{\theta_1\sum_{j=0}^{r}\binom{r}{j}\mu_{r-j,s}(\lambda^*+1)+\theta_3\sum_{j=0}^{r}\sum_{k=0}^{s}\binom{r}{j}\binom{s}{k}\mu_{r-j,s-k}(\lambda^*+1)\} \tag{3.18}$$

Now equate the coefficients of $(r!s!)^{-1}(it_1)^r(it_2)^s$ on both sides of (3.18) to get the relation (3.15). A similar procedure gives (3.16). □

4 Estimation of parameters

Here we obtain the estimators of the $BHPD$ by the method of maximum likelihood. Let $a(r,s)$ be the observed frequency of the $(r,s)^{th}$ cell of the bivariate data. Let y be the highest value of r observed and z be the highest value of s observed. Then by using (2.11) the likelihood function of the sample is the following.

$$L=\prod_{r=0}^{y}\prod_{s=0}^{z}[h(r,s)]^{a(r,s)} \quad\Rightarrow\quad \log L=\sum_{r=0}^{y}\sum_{s=0}^{z}a(r,s)\log h(r,s).$$

Let $\hat\theta_1$, $\hat\theta_2$, $\hat\theta_3$ and $\hat\lambda$ denotes the likelihood estimators of θ_1, θ_2, θ_3 and λ respectively. Now $\hat\theta_1$, $\hat\theta_2$, $\hat\theta_3$ and $\hat\lambda$ are obtained by solving the likelihood equations (4.1), (4.2), (4.3)and (4.4) given below.

$$\frac{\partial\log L}{\partial\theta_1}=0$$

Equivalently,

$$\sum_{r=0}^{y}\sum_{s=0}^{z}a(r,s)\left\{\frac{-1}{\lambda}\frac{\phi(2;\lambda+1;\theta_1+\theta_2+\theta_3)}{\phi(1;\lambda;\theta_1+\theta_2+\theta_3)}+\frac{\sum_{m=0}^{\min(r,s)}\frac{D^*\theta_1^{r-m-1}\theta_2^{s-m}\theta_3^m}{(r-m-1)!(s-m)!m!}}{\xi(r,s)}\right\}=0. \tag{4.1}$$

$$\frac{\partial\log L}{\partial\theta_2}=0$$

Equivalently,

$$\sum_{r=0}^{y}\sum_{s=0}^{z}a(r,s)\left\{\frac{-1}{\lambda}\frac{\phi(2;\lambda+1;\theta_1+\theta_2+\theta_3)}{\phi(1;\lambda;\theta_1+\theta_2+\theta_3)}+\frac{\sum_{m=0}^{\min(r,s)}\frac{D^*\theta_1^{r-m}\theta_2^{s-m-1}\theta_3^m}{(r-m)!(s-m-1)!m!}}{\xi(r,s)}\right\}=0. \tag{4.2}$$

$$\frac{\partial \log L}{\partial \theta_3} = 0$$

Equivalently,

$$\sum_{r=0}^{y}\sum_{s=0}^{z} a(r,s)\left\{\frac{-1}{\lambda}\frac{\phi(2;\lambda+1;\theta_1+\theta_2+\theta_3)}{\phi(1;\lambda;\theta_1+\theta_2+\theta_3)} + \frac{\sum_{m=0}^{\min(r,s)} \frac{D^*\theta_1^{r-m}\theta_2^{s-m}\theta_3^{m-1}}{(r-m)!(s-m)!(m-1)!}}{\xi(r,s)}\right\} = 0. \tag{4.3}$$

$$\frac{\partial \log L}{\partial \lambda} = 0$$

Equivalently,

$$\sum_{r=0}^{y}\sum_{s=0}^{z} a(r,s)\left\{\frac{-1}{\phi(1;\lambda;\theta_1+\theta_2+\theta_3)}\sum_{x=0}^{\infty}(\theta_1+\theta_2+\theta_3)^x\eta(x)+\right.$$

$$\left.\frac{1}{\xi(r,s)}\sum_{m=0}^{\min(r,s)}\eta(r+s-m)\,\frac{(r+s-m)!\theta_1^{r-m}\theta_2^{s-m}\theta_3^{m}}{(r-m)!(s-m)!m!}\right\} = 0, \tag{4.4}$$

in which $\xi(r,s) = \sum_{m=0}^{\min(r,s)} \frac{D^*\theta_1^{r-m}\theta_2^{s-m}\theta_3^{m}}{(r-m)!(s-m)!m!}$ and $\eta(u) = \frac{\Gamma(\lambda)}{\Gamma(\lambda+u)}[\psi(\lambda) - \psi(\lambda+u)]$.

5 An application

Here we illustrate the method of maximum likelihood estimation using a real life data set taken from Patrat (1993). The description of data is as follows: The North Atlantic coastal states in USA can be affected by tropical cyclones. They divided the states into three geographical zones: Zone 1 (Texas, Louisina, The Mississipi, Alabama), Zone 2 (Florida), and Zone 3 (Other states)

Now the interest is in the study of the joint distribution of the pair (X_1, X_2), where X_1 and X_2 are the yearly frequency of hurricanes affecting respectively zone 1 and zone 3. The observed values of (X_1, X_2) during 93 years from 1899 to 1991 are as given in Table 1. We obtain the corresponding expected frequencies by fitting the bivariate Poisson distribution (BPD), the bivariate hyper-Poisson distribution of Ahmad (1981) ($BHPD_A$) and the bivariate hyper-Poisson distribution ($BHPD$) introduced in this paper using method of maximum likelihood in Table 1. The estimated values of the parameters of the BPD, the $BHPD_A$ and the $BHPD$ and the chi-square values in respective cases are listed in Table 2. From Table 2, it can be observed that the $BHPD$ gives a better fit to this data compared to the existing models- the BPD and the $BHPD_A$.

Table 1: Comparison of observed and theoretical frequencies Hurricanes (1899-1991) having affected Zone 1 and Zone 3, using method of maximum likelihood.

Zone 1 / Zone 3		0	1	2	3	Total
0	OBS	27	9	3	2	41
	BPD	28.24	12.71	2.86	0.48	44.29
	$BHPD_A$	28.31	12.49	2.95	0.48	44.23
	$BHPD$	25.64	14.31	2.50	0.26	42.71
1	OBS	24	13	1	0	38
	BPD	20.30	9.79	2.35	0.42	32.86
	$BHPD_A$	20.46	9.56	2.37	0.40	32.79
	$BHPD$	23.23	10.88	2.23	0.27	36.61
2	OBS	8	2	1	0	11
	BPD	7.29	3.75	0.96	0.19	12.19
	$BHPD_A$	7.39	3.65	0.95	0.17	12.16
	$BHPD$	6.60	3.62	0.88	0.12	11.22
3	OBS	1	0	2	0	3
	BPD	2.12	1.16	0.32	0.06	3.66
	$BHPD_A$	1.78	0.93	0.25	0.05	3.01
	$BHPD$	1.11	0.72	0.21	0.14	1.07
Total	OBS	60	24	7	2	93
	BPD	57.95	27.41	6.49	1.15	93
	$BHPD_A$	57.94	26.63	6.52	1.1	92
	$BHPD$	56.58	29.53	5.82	0.79	93

Table 2: Estimated values of the parameters of the BPD, the $BHPD_A$ and the $BHPD$ by the method of maximum likelihood estimation and corresponding chi-square values.

Distributions	Estimation of parameters	Chi−square values
BPD	$\hat{\theta}_1 = 0.683, \hat{\theta}_2 = 0.450, \hat{\theta}_3 = 0.021$	2.524
$BHPD_A$	$\hat{\theta}_1 = 0.780, \hat{\theta}_2 = 0.324, \hat{\theta}_3 = 0.021$	2.452
	$\hat{\lambda}_1 = 1.075, \hat{\lambda}_2 = 0.619$	
$BHPD$	$\hat{\theta}_1 = 0.414, \hat{\theta}_2 = 0.255, \hat{\theta}_3 = 0.049$	0.463
	$\hat{\lambda} = 0.457$	

References

[1] Ahmad, M. (1979). A note on the moments of hyper-Poisson distribution. *Arabian Journal of Science and Engineering*, **4**, 65–68.

[2] Ahmad, M. (1981). On a bivariate hyper-Poisson distribution. Statistical Distributions in scientific work 4. G.P. Patil, S. Kotz, J.K. Ord, (editors) 225–230. Dordrecht, Reidel.

[3] Ahmad, M. (2007). A short note on Conway-Maxwell-hyper Poisson distribution. *Pakistan Journal of Statistics*, **23**, 135–137.

[4] Bardwell, G. E. and Crow, E. L. (1964). A two parameter family of hyper-Poisson distribution. *Journal of American Statistical Association* **59**, 133–141.

[5] Crow, E. L. and Bardwell, G. E. (1965). Estimation of the parameters of the hyper-Poisson distributions. *In classical and contagious discrete distributions* G. P. Patil, (editor), 127–140.

[6] Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). Univariate Discrete Distributions, 3^{rd} Edition Wiley, New York.

[7] Kemp, C. D. (2002). q-analogues of the hyper-Poisson distribution. *Journal of Statistical Planning and Inference* **101**, 179–183.

[8] Kocherlakotta, S. and Kocherlakotta, K. (1992). *Bivariate discrete distributions*. Marcel Dekker, New York.

[9] Kumar, C. S. (2008). A unified approach to bivariate discrete distributions. *Metrika* **67**, 113–123.

[10] Kumar, C. S. and Nair, B. U. (2011). A Modified version of hyper-Poisson distribution and its applications. *Journal of Statistics and Applications* **6**, 25–36.

[11] Kumar, C. S. and Nair, B. U. (2012a). An extended version of hyper-Poisson distribution and some of its applications. *Journal of Applied Statistical Sciences* **19**, 81–88.

[12] Kumar, C. S. and Nair, B. U. (2012b). An Alternative hyper-Poisson distribution. *Statistica* **72**, 357–369.

[13] Mathai, A. M. and Saxena, R. K. (1973). *Generalised Hypergeometric Functions with Applications in Statistics and Physical sciences.* Springer-Verlag, Heidelberg.

[14] Nisida, T. (1962). On the multiple exponential channel queuing system with hyper-Poisson arrivals. *Journal of the Operations Research Society* **5**, 57–66.

[15] Patrat, C. (1993). Compound model for two dependent kinds of claim. *ASTIN Colloquium*, XXIVe Cambridge.

[16] Roohi, A. and Ahmad, M. (2003a). Estimation of the parameter of hyper-Poisson distribution using negative moments. *Pakistan Journal of Statistics* **19**, 99–105.

[17] Slater, L. J. (1960). *Confluent hypergeometric functions*. Cambridge University Press, Cambridge.

[18] Staff, P. J. (1964). The displaced Poisson distribution. *Australian Journal of Statistics* **6**, 12–20.

[19] Subrahmaniam, K. (1966). A test of for intrinsic correlation in the theory of accident proneness. *Journal of the Royal Statistical Society, Series B* **28**, 180–189.

Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 1, pp. 47-59

DEFINITIVE TESTING OF AN INTEREST PARAMETER: USING PARAMETER CONTINUITY

D. A. S. FRASER

Department of Statistical Sciences, University of Toronto, Toronto, Canada M5S 3G3, Canada
Email: dfraser@utstat.toronto.edu

SUMMARY

For a scalar or vector parameter of interest with a regular statistical model, we determine the definitive null density for testing a particular value of the interest parameter: continuity gives uniqueness without reference to sufficiency but the use of full available information is presumed. We start with an exponential family model, that may be either the original model or an approximation to it obtained by ancillary conditioning. If the parameter of interest is linear in the canonical parameter, then the null density is third order equivalent to the conditional density given the nuisance parameter score; and when the parameter of interest is also scalar then this conditional density is the familiar density used to construct unbiased tests. More generally but with scalar parameter of interest, linear or curved, this null density has distribution function that is third order equivalent to the familiar higher-order p-value $\Phi(r^*)$. Connections to the bootstrap are described: the continuity-based ancillary of the null density is the natural invariant of the bootstrap procedure. Also ancillarity provides a widely available general replacement for the sufficiency reduction. Illustrative examples are recorded and various further examples are available in Davison et al. (2014) and Fraser et al. (2016).

Keywords and phrases: Ancillary; Exponential model; Information; Likelihood asymptotics; Nuisance parameter; p-value; Profile likelihood; Score conditioning; Similar test

1 Introduction

We consider the problem of testing a value for a d-dimensional parameter of interest ψ in the presence of a $(p-d)$-dimensional nuisance parameter λ, in the context of a statistical model $f(y;\psi;\lambda)$ on $\mathbb{R}^n$ that we assume has the usual regularity conditions for deriving higher order approximations. We show that continuity and ancillarity directly determine a density that is free of the nuisance parameters, a density that can be viewed as providing measurement of the parameter of interest. The saddlepoint approximation then gives an expression for this density with error of $O(n^{-3/2})$. If the parameter of interest is scalar, inference based on this null density leads immediately to the familiar r^* approximation (Barndorff-Nielsen, 1991; Fraser, 1990; Brazzale et al., 2007). An associated average p-value can also be approximated to the same order by a parametric bootstrap, as initiated in

Lee and Young (2005), Fraser and Rousseau (2008) and DiCiccio and Young (2008); computation time and ease of use can however differ dramatically.

In §2 we present the model, in §3 develop the null density 3.4 for testing the interest parameter ψ, and then in §4 specialize this to the linear interest parameter case obtaining the the null density 4.3; this is then shown to be equivalent to the familiar conditional distribution 4.5, which in the scalar interest case is widely used to derive unbiased or similar tests. In §5, for ψ a scalar parameter, we relate the null density to the higher-order likelihood based p-values obtained from the familiar r^* approximation. For a vector ψ we propose the use of directional p-values, which can be obtained by one-dimensional integration. Numerical examples of the latter application are given in Davison et al. (2014) and Fraser et al. (2016). Intrinsic connections with the parametric bootstrap are addressed in §6.

2 Exponential model

Suppose we have a statistical model $f(y;\theta)$ for a response $y \in \mathbb{R}^n$ with parameter $\theta \in \mathbb{R}^p$ that takes the exponential form

$$f(y;\theta) = \exp\left[\varphi(\theta)^\tau v(y) - \kappa\{\varphi(\theta)\}\right]h(y), \tag{2.1}$$

where the canonical $\varphi(\theta)$ in $\mathbb{R}^p$ is one-to-one equivalent to θ, and the canonical $v(y)$ in $\mathbb{R}^p$ is the usual variable directly affected by the parameter. The assumption of exponential form is more general than it may appear, as this form arises widely with regular statistical models as the tangent exponential approximation, tangent at the observed value y^0 with tangent vectors V. The construction of the tangent exponential model is briefly outlined in Appendix A, together with references to the literature.

Two key simplifications offered by 2.1 are that the distribution of v provides all the information about φ and that the density of v can be approximated by the saddlepoint method. Thus our model for inference can be written

$$g(v;\varphi) = \exp\left\{\varphi^\tau v - \kappa(\varphi)\right\}g(v) \tag{2.2}$$

$$= \frac{e^{k/n}}{(2\pi)^{p/2}} \exp\left\{\ell(\varphi;v) - \ell(\hat\varphi;v)\right\}|\jmath_{\varphi\varphi}(\hat\varphi)|^{-1/2}\{1 + O(n^{-3/2})\}, \tag{2.3}$$

where $\ell(\varphi;v) = \varphi^\tau v - \kappa(\varphi)$ is the log-likelihood function, $\hat\varphi = \hat\varphi(v)$ is the maximum likelihood estimator, $\jmath_{\varphi\varphi}(\hat\varphi) = -\partial\ell/\partial_\varphi\partial_\varphi^T|_{\hat\varphi}$ is the observed information array in the canonical parameterization, and k/n is a generic normalizing constant (Daniels, 1954; Barndorff-Nielsen and Cox, 1979). From some original regular model this approximation needs only the observed log-likelihood function $\ell^0(\theta)$ from y^0 and the observed gradient $\varphi(\theta)$ of the log-likelihood in the directions V, and then effectively implements the integration for the original model or its approximation 2.1 to produce the marginal density $g(v;\varphi)$ to third order from that of y.

3 Curved interest and exponential model

In 2.2 and 2.3 we suppressed the dependence of φ on θ for convenience; and we now assume that our parameter of interest is $\psi(\varphi) \in \mathbb{R}^d$, and use 2.3 to obtain the density 3.4 for testing $\psi(\varphi) = \psi_0$, eliminating the nuisance parameter λ. Thus, we consider $\psi(\varphi)$ to be fixed at ψ_0 in 2.3, so the model has a p-dimensional variable v, and a $(p-d)$-dimensional unknown parameter λ. With $\psi(\varphi)$ fixed at ψ_0, there is an approximate ancillary statistic S for λ, a function of v with a marginal distribution free of λ (Fraser et al., 2010), and the ancillary density is uniquely determined to $O(n^{-3/2})$. Thus the reference marginal density for inference about a value ψ based on this function of v is also unique.

To describe this density we define a plane L^0 in the sample space by fixing the constrained maximum likelihood estimator of λ at its observed value:

$$L^0 = \{v \in \mathbb{R}^p : \hat{\lambda}_{\psi_0} = \hat{\lambda}^0_{\psi_0}\}$$

where $\hat{\lambda}_{\psi_0}(v)$ is obtained as the solution of the score equation $\partial\ell(\varphi; v)/\partial\lambda = 0$ with notation $\hat{\lambda}_{\psi_0}(v^0) = \hat{\lambda}^0_{\psi_0} = \bar{\lambda}^0$. The constrained estimate of the full parameter φ at (s, t^0) is $\bar{\varphi}^0$. In some generality the interest parameter ψ can be non-linear; in that case we define a new parameter $\chi = \chi(\varphi)$ linear in φ that is tangent to $\psi(\varphi)$ at $\bar{\varphi}^0$; the right hand panel of Figure 1 shows the curve with ψ fixed, the constrained maximum likelihood estimate $\bar{\varphi}^0$, and the linear approximation

$$\chi(\varphi) = \psi(\bar{\varphi}^0) + \bar{\psi}^0_\varphi(\varphi - \bar{\varphi}^0), \tag{3.1}$$

as well as the overall maximum likelihood estimate $\hat{\varphi}^0$; here $\bar{\psi}^0_\varphi = (\partial\psi/\partial\varphi)|_{\bar{\varphi}^0}$ is the needed Jacobian. The complementing parameter λ in the full parameter space is shown in Figure 1 as orthogonal to χ, for convenience. The left panel of Figure 1 shows the sample space, using corresponding rotated canonical variables s and t: in particular the profile plane L^0 on the sample space corresponds to a $p-d$ dimensional variable t, fixed at its observed value t^0. The d-dimensional variable s on L^0 indexes the ancillary contours where they intersect L^0. In effect (s, t) plays the role of the full canonical variable in an approximating exponential model, and χ is linear in the canonical parameter.

On L^0 the saddlepoint approximation to the joint density is, from 2.3

$$g(s, t^0) = \frac{e^{k/n}}{(2\pi)^{p/2}} \exp\{\ell(\bar{\varphi}; s, t^0) - \ell(\hat{\varphi}; s, t^0)\}|j_{\varphi\varphi}(\hat{\varphi})|^{-1/2}, \tag{3.2}$$

where $\hat{\varphi} = \hat{\varphi}(s, t^0)$. The conditional density of t given the ancillary labelled by $S = s$ has a $p^\star$ approximation at its maximum which when evaluated on L^0 at $\bar{\varphi}$ simplifies to

$$\frac{e^{k/n}}{(2\pi)^{(p-d)/2}}|j_{(\lambda\lambda)}(\bar{\varphi})|^{-1/2}. \tag{3.3}$$

The marginal density for the ancillary variable S as indexed by s on the observed L^0 is then obtained by dividing the joint density 3.2 at $(s; t_0)$ by the conditional density 3.3 of t given the

ancillary S, with both evaluated at (s, t^0) on L^0:

$$g_m(s;\psi_0) = \frac{e^{k/n}}{(2\pi)^{d/2}} \exp\{\ell(\bar\varphi; s, t^0) - \ell(\hat\varphi; s, t^0)\} |\jmath_{\varphi\varphi}(\hat\varphi)|^{-1/2} |\jmath_{(\lambda\lambda)}(\bar\varphi|^{1/2} ds, \tag{3.4}$$

to third order. In 3.4, the exponent $\ell(\bar\varphi) - \ell(\hat\varphi)$ is the log-likelihood ratio statistic at (s, t^0) for the tested value ψ_0, and the nuisance information determinant in the exponential parameterization (λ) can be obtained from that in terms of λ by applying the Jacobian φ_λ,

$$|\jmath_{(\lambda\lambda)}(\psi_0, \hat\lambda^0_{\psi_0})| = |\jmath_{\lambda\lambda}(\bar\varphi^0)|\, |\varphi^T_\lambda(\bar\varphi^0)\varphi_\lambda(\bar\varphi^0)|^{-1}, \tag{3.5}$$

as described in Fraser and Reid (1993), Brazzale et al. (2007) or Davison et al. (2014). In the left panel of Figure 1 we show the curve $\psi(\varphi) = \psi_0$, and two different lines L^0 and L^{00} corresponding to two different points u^0 and u^{00} on an ancillary contour for the particular ψ_0 value.

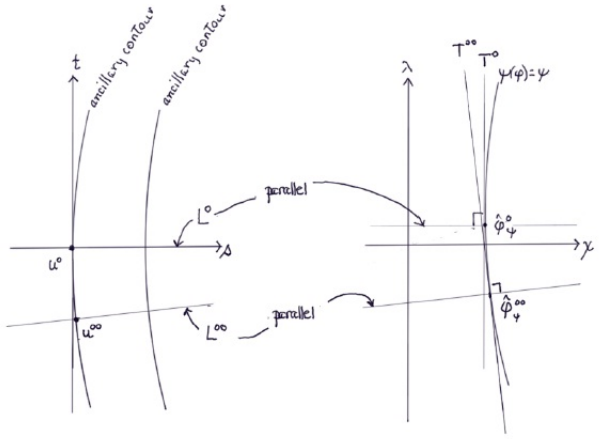

Figure 1: Score space on left; canonical parameter space on right; ancillary contours through observed u^0 and through a nearby point u^{00} on same ancillary contour for ψ.

The ancillary distribution 3.4 for testing is recorded in terms of s on L^0 but represents the result of integrating along the ancillary contours relative to $\psi(\varphi) = \psi_0$, not by integrating for fixed s; accordingly the distribution appears to depend on t^0, but this is an artifact of its presentation using coordinates that do depend on t^0 (Fraser and Reid, 1995; Fraser and Rousseau, 2008); see Example 4.1 in the next section. The ancillary distribution is developed above within an exponential model, either the given model or a tangent approximation to it as described in Appendix A. The development for a regular model from the point of view of approximate studentization is available in Fraser and Rousseau (2008); the distribution has third order uniqueness even though the third order ancillary itself is not unique.

4 Linear interest and exponential model

Consider a special case of the exponential model 2.1 where the interest parameter $\psi = \chi$ is linear and the full canonical parameter φ is just (φ, λ):

$$g(\upsilon;\theta) = \exp\{\upsilon_1^T\psi + \upsilon_2^T\lambda - \kappa(\psi,\lambda)\}h(\upsilon). \tag{4.1}$$

It is helpful to centre υ at the observed value: letting $s = \upsilon_1 - \upsilon_1^0$ and $t = \upsilon_2 - \upsilon_2^0$ gives

$$g(s,t;\theta) = \exp\{s^T\psi + t^T\lambda + \ell^0(\psi,\lambda)\}h(s,t). \tag{4.2}$$

where $\ell^0(\psi,\lambda)$ is the negative cumulant generating function for the latent density $h(s,t)$. The marginal density 3.4 then simplifies to

$$g_m(s;\psi) = \frac{e^{k/n}}{(2\pi)^{d/2}}\exp\{\ell(\hat\theta_\psi) - \ell(\hat\theta)\}|\jmath_{\theta\theta}(\hat\theta)|^{-1/2}|\jmath_{\lambda\lambda}(\hat\theta_\psi)|^{1/2}ds, \tag{4.3}$$

where $\hat\theta = \hat\theta(s,t^0)$ and $\hat\theta_\psi = (\psi, \hat\lambda^0_\psi)$, all to third order.

The conditional density of s given t is more conventionally used for inference about ψ in this linear setting. From 4.2 we have

$$g_c(s|t;\psi) = \exp\{s^T\psi - \kappa_t(\psi)\}h_t(s), \tag{4.4}$$

and its saddlepoint approximation is

$$g_c(s|t;\psi) = \frac{e^{k/n}}{(2\pi)^{d/2}}\exp\{\ell_P(\psi) - \ell_P(\hat\psi)\}|\jmath_P(\hat\psi)|^{-1/2}\left\{\frac{|\jmath_{\lambda\lambda}(\psi,\hat\lambda^0_\psi)|}{|\jmath_{\lambda\lambda}(\hat\psi,\hat\lambda)|}\right\}^{1/2}, \tag{4.5}$$

where $\ell_P(\psi)$ is the profile log-likelihood function $\ell(\psi,\hat\lambda_\psi)$, and $\jmath_P(\psi) = -\partial^2\ell_P(\psi)/\partial\psi\partial\psi^T$ is the associated information function (Davison, 1988; Barndorff-Nielsen and Cox, 1979). The two densities 4.3 and 4.5 are identical, as $|\jmath_{\theta\theta}(\hat\theta)| = |\jmath_P(\hat\psi)||\jmath_{\lambda\lambda}(\hat\theta)|$.

In the above we have not distinguished the tested value ψ_0, because in the exponential model with canonical interest parameter ψ, the planes L^0 for different tested values of ψ are parallel, and the distributions as recorded on each plane are equivalent, so the resulting marginal density 4.3 can be used as the pivotal quantity to test any value of ψ and thereby provide confidence intervals or regions.

Example 4.1. We illustrate this with a simple exponential model, for which the detailed calculations are readily obtained. Take $p = 2$ and suppose that the joint density of s, t is of the form

$$g(s,t;\psi,\lambda) = \phi(s-\psi)\phi(t-\psi)\exp\{-a\psi\lambda^2/(2n^{1/2})\}h(s,t), \tag{4.6}$$

where $\phi(.)$ is the standard normal density. The function $h(s,t)$ can be explicitly obtained as

$$h(s,t) = 1 + \frac{1}{2}as(t^2-1)n^{-1/2} + \frac{1}{8}a^2(s^2-1)(t^4-6t^2+3)n^{-1} + O(n^{-3/2}), \tag{4.7}$$

and we can re-write the density as

$$g(s,t;\psi,\lambda) = \{1 - a\psi\lambda^2/(2n^{1/2}) + a^2\psi^2\lambda^4/(8n)\}\phi(s-\psi)\phi(t-\psi) \times \{1 + as(t^2-1)/(2n^{1/2}) + a^2(s^2-1)(t^4-6t^2+3)/(8n) + O(n^{-3/2})\}. \quad (4.8)$$

The related marginal density is obtained by taking all terms in 4.8 to the exponent and completing the square; this shows that, ignoring terms of $O(n^{-3/2})$, there is a pivotal function Z_ψ which follows a standard normal distribution to third order:

$$Z_\psi = s\{1 + a^2(2t^2-1)/4n\} - \psi\{1 - a^2(2t^2-1)/4n\} - a(t^2-1)/2n^{1/2}. \quad (4.9)$$

From this we see that s has conditional bias $a(t^2-1)/2n^{1/2} + O(n^{-1})$, but this bias in the measurement of ψ is of no consequence for inference, as it is removed as part of forming the pivot Z_ψ. If we ignore terms of $O(n^{-1})$ then $s - a(t^2-1)/(2n^{1/2})$ is standard normal to $O(n^{-1})$, *i.e.* to this order only a location adjustment is needed to obtain an approximately standard normal pivotal quantity.

5 Inference for ψ from the reference density

The base density $g_m(s;\psi)$ on $\mathbb{R}^d$ given at 3.4 is to third order the unique density for inference about ψ, in the sense that it is a direct consequence of requiring model continuity to be retained in the elimination of the nuisance parameter (Fraser et al., 2010). The density can be computed from the distribution 2.2 or 2.3 for the canonical variable u or from the observed log-likelihood from the original model $\ell(\varphi;y^0) = \log\{f(y^0;\varphi)\}$ together with the observed log-likelihood gradient $\varphi(\theta) = \ell_{;V}(\theta;y^0)$ in directions V; see Appendix A.

If $d = 1$, the one-dimensional density can be integrated numerically. It can also be shown to be third-order equivalent to a standard normal density for the familiar pivot $r^* = r^*(\psi;y^0)$, defined by

$$r^*(\psi;y^0) = r - r^{-1}\log\frac{r}{Q}, \quad (5.1)$$

$$r = \pm\left(2[\ell\{\varphi(\hat\theta);y^0\} - \ell\{\varphi(\hat\theta_\psi);y^0\}]\right)^{1/2}, \quad (5.2)$$

$$Q = \pm|\chi(\hat\theta) - \chi(\hat\theta_\psi)|/\hat\sigma_\chi, \quad (5.3)$$

where $\hat\sigma_\chi^2 = |\jmath_{(\lambda\lambda)}\{\varphi(\hat\theta_\psi)\}|/|\jmath_{\varphi\varphi}\{\varphi(\hat\theta)\}|$ is a particular estimate of the variance of the numerator of Q, and $\pm$ designates the sign of $\hat\psi^0 - \psi$. From the definition 3.1 of χ as tangent to ψ at $\varphi(\hat\theta_\psi)$, we obtain an alternate expression for Q,

$$Q = \frac{|\varphi(\hat\theta) - \varphi(\hat\theta_\psi)\ \varphi_\lambda(\hat\theta_\psi)|}{|\varphi_\theta(\hat\theta)|}\,\frac{|\jmath_{\theta\theta}(\hat\theta)|^{1/2}}{|\jmath_{\lambda\lambda}(\hat\theta_\psi)|^{1/2}} \quad (5.4)$$

which can be more convenient for computation. Several examples of the use of $r^*(\psi;y^0)$ as a standard normal pivotal for inference about a scalar parameter of interest are given in Fraser et al. (1999) and Brazzale et al. (2007). For Example 4.1, straightforward calculations verify that $r^*(\psi;y^0) = Z_\psi$ of 5.4 to $O(n^{-3/2})$.

Example 5.1. As an illustration of exact and approximate p-value contours, consider two exponential life variables y_1, y_2 with failure rates φ_1, φ_2 and with interest parameter chosen as the total failure rate $\psi = \varphi_1 + \varphi_2$; the model is $\varphi_1\varphi_2 \exp\{-\varphi_1 y_1 - \varphi_2 y_2\}$ with $0 < y_1, y_2 < \infty$. A rotation of variable and of parameter through $\pi/4$ gives new variables $s = (y_1+y_2)/2^{1/2}, t = (-y_1+y_2)/2^{1/2}$ and new parameters $\chi = (\varphi_1 + \varphi_2)/2^{1/2}, \lambda = (-\varphi_1 + \varphi_2)/2^{1/2}$ on equivalent rotated quadrants; the model then becomes

$$f(s,t) = (\chi^2/2 - \lambda^2/2)\exp(-\chi s - \lambda|t|),$$

with $s > |t| > 0, \chi > |\lambda| > 0$ and parameter of interest $\psi = 2^{1/2}\chi$. The exact conditional density of s, given t, is $f(s|t;\chi) = \chi \exp\{-\chi(s-|t|)\}$, for $s > |t|$, i.e. the pivotal quantity $Z_\chi = \chi(s-|t|)$ follows an exponential distribution with rate 1. The approximation 4.3 is an $O(n^{-3/2})$ approximation to this, equivalent to a standard normal distribution for the adjusted log-likelihood root r^*_χ.

In Figure 5 we illustrate three quantile contours, at levels 25%; 50%, 75%, for the exact conditional distribution and for the normal approximation to the distribution of r^*, for testing the value of $\psi = 0.6931$ or equivalently $\chi = 0.4901$. The contours of the exact conditional density are line segments, and the contours of the approximate normal distribution for r^*_χ are smooth curves. The conditional and marginal approaches are identical to third order: the difference that appears in Figure 5 is due entirely to the approximation to the marginal density. From one point of view the normal approximation to r^*_χ replaces exact similarity of the test with similarity to $O(n^{-3/2})$, and the smoothed version is somewhat less sensitive to the exact value of t. Third-order similarity of tests based on r^* is established in Jensen (1992).

Example 5.2. As an illustration of an exponential model with a curved interest parameter suppose in Example 5.1 that the parameter of interest is now taken to be $\psi = \varphi_1\varphi_2$; we let $\lambda = \varphi_2$ be an initial nuisance parameter. Then

$$\hat\psi = 1/(y_1y_2), \quad \hat\lambda = 1/y_1, \quad \hat\lambda^2_\psi = \psi_1/y_2 = \psi\hat\lambda^2/\hat\psi.$$

The linear parameter $\chi(\varphi)$ is

$$\chi(\varphi) = \chi(\bar\varphi) + \psi_\varphi(\bar\varphi)(\varphi - \bar\varphi) = \chi(\bar\varphi) - 2\bar\varphi_1\bar\varphi_2 + \bar\varphi_2\varphi_1 + \bar\varphi_1\varphi_2$$

and s is the corresponding linear combination of y_1 and y_2. The information determinants $|j_{\varphi\varphi}(\hat\varphi)|$ and $|j_{(\lambda\lambda)}(\bar\varphi)|$ are $\hat\psi^{-2}$ and $2\psi\hat\lambda_\psi\hat\lambda/(\hat\psi(\psi^2+1))$, respectively; the latter is obtained using $\varphi_\lambda = (-\psi/\lambda^2, 1)$. The marginal density of s is approximated by 3.4, from which it follows that r^*_ψ is a pivotal quantity following a standard normal distribution to $O(n^{-3/2})$. The quantile curves will be similar in shape to those in Example 4.1, but there is no exact conditional density for comparison.

When ψ is a vector, the marginal density on L^0 does not immediately lead to a single p-value function. A directional approach is available following Fraser and Massam (1985), Skovgaard (1988), Fraser and Reid (2006) and references therein. On L^0 the mean value under the fitted null

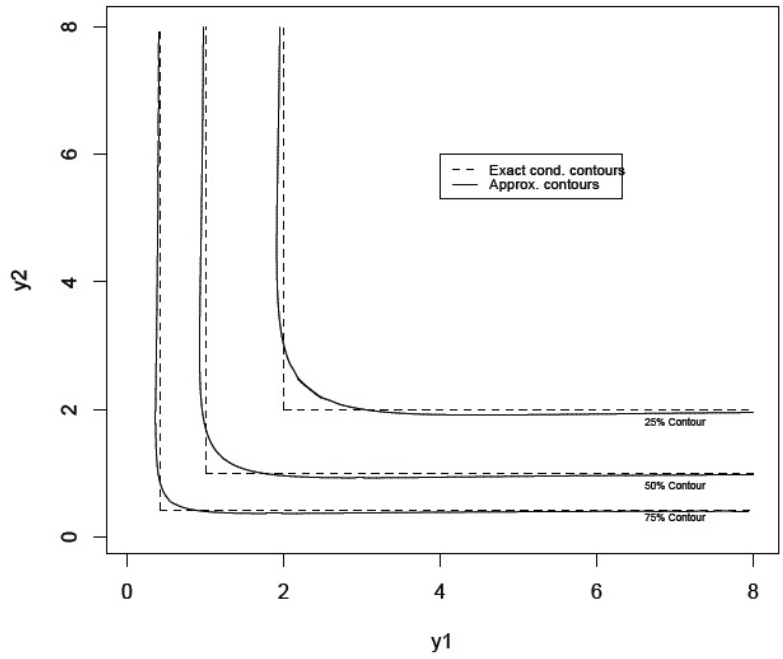

Figure 2: The exact conditional contours and the third order approximate contours at quantile levels 25%, 50% and 75% for testing $\psi = 0.6931$ in the simple exponential life model.

parameter value $\hat{\varphi}^0_\psi$ is $s_\psi = -\ell^0_\psi(\psi, \hat{\lambda}^0_\psi)$ with corresponding data $s = 0$ in the standardized coordinates. Then conditioning on the direction from expected s_ψ to observed $s = 0$ gives the directional p-value

$$p(\psi) = \frac{\int_1^\infty g_m\{s_\psi + t(0 - s_\psi)\}t^{d-1}dt}{\int_0^\infty g_m\{s_\psi + t(0 - s_\psi)\}t^{d-1}dt};$$

which can easily be evaluated numerically. A number of examples based on familiar exponential models where calculations are particularly accessible are presented in Davison et al. (2014) and Fraser et al. (2016).

6 Bootstrap and higher order likelihood

For the exponential model 2.1, improved p-values for testing a scalar interest parameter $\psi = \psi_0$ can also be obtained using bootstrap sampling from the estimated model $f(y; \hat{\theta}^0_\psi)$. In particular, this bootstrap applied to the signed log-likelihood root $r_\psi = \pm[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2}$ produces p-values

that are uniformly distributed with accuracy $O(n^{-3/2})$, and are asymptotically equivalent to this order to p-values obtained from the normal approximation to r^*_ψ. If however the estimated sampling model is taken to be the traditional $f(y;\hat\theta^0)$ then the relative error drops to $O(n^{-1})$ (DiCiccio et al (2001); Lee & Young (2005)).

The bootstrap has an intrinsic connection with the ancillary distribution 3.3 for the marginal variable S, as recorded in terms of s on the observed L^0. Indeed, the bootstrap distribution $f(y;\hat\theta^0_\psi)$ directly produces the preceding null distribution for the ancillary S; this follows by noting that the distribution of S is free of λ and thus a particular choice $\lambda = \hat\lambda^0_\psi$ in the re-sampling model just generates the same marginal null distribution. Accordingly the distribution 3.3 can be viewed as an invariant of the bootstrap procedure. It also follows that the bootstrap distribution of any statistic that is a function of the ancillary S is also an invariant of the bootstrap distribution to third order.

More generally with an asymptotic model having full parameter dimension p, and null parameter dimension $p-d$, the moderate deviations region can be presented as a product space with coordinates $(S,\hat\lambda_\psi)$ and a bootstrap step can be viewed as a projection along contours of the ancillary variable S such that dependence on the conditional $\hat\lambda_\psi$ is reduced by a factor $O(n^{-1/2})$ (Fraser and Rousseau, 2008).

Meanwhile for the higher-order likelihood approach, the standard normal approximation to the usual r^*_ψ is accurate to $O(n^{-3/2})$. This, with the preceding bootstrap result, shows that the higher-order r^*_ψ approximation can be implemented directly by bootstrap resampling of r^*_ψ or equivalently the bootstrap resampling of r_ψ which is known to be affinely equivalent to r^*_ψ to third order, using of course the estimated null model $f(y;\hat\theta^0_\psi)$; computation times however can be significantly different: for a recent example calculations used 20 hours for the bootstrap calculation to achieve the same accuracy as the higher order likelihood calculation achieved in 0.09 seconds.

For an exponential model with scalar linear interest parameter, DiCiccio & Young (2008) show that the null model bootstrap distribution of r^*_ψ directly approximates the conditional distribution of r^*_ψ even though the bootstrap is an unconditional simulation; this follows from 4.3 and 4.5 by noting that the marginal and conditional distributions are identical to third order.

More generally with a regular model and conditioning based on tangent vectors V a bootstrap step provides an average over the conditioning indexed by the vectors V and thus does not record the precision information that is routinely available by the higher order approach and even certain higher order Bayesian calculations. Thus we can say that the parametric bootstrap based on the observed maximum likelihood estimate under the null reproduces an average of the higher order r^*_ψ evaluations rather than the individual precision-tuned p-values coming from the higher-order method.

7 Conclusion

For general regular models with scalar or vector, linear or curved interest parameters, we have determined the score space distribution that has nuisance parameter effects removed and has the full information to provide r^* tests for scalar parameters and directed tests for vector parameters. We have thus extended available distribution theory for statistical inference, and integrated the direc-

tional methodology with the higher order distribution theory. In particular for the vector parameter case this can fine-tune the Bartlett-corrected l-dimensional numerical integration (Davison et al., 2014).

Acknowledgment

This research is part of a continuing joint research project on statistical inference using methods of likelihood asymptotics, joint with Nancy Reid, Department of Statistical Sciences, University of Toronto, and received support from the Natural Sciences and Engineering Research Council of Canada, the Department of Statistical Sciences at the University of Toronto, and the Senior Scholars Funding of York University. Very special thanks to Mahbub Latif for support with type setting, and to Wei Lin and Uyen Hoang for manuscript development.

Appendix A

In the context of a regular statistical model with continuity suppose that the variable y has n independent coordinates, with $n > p$, the dimension of the parameter θ. Let $p = p(y;\theta)$ be the vector with ith coordinate $p_i = F_i(y_i;\theta)$, where $F_i(.;\theta)$ is the distribution function for the ith component of y. By inverting $p = p(y;\theta)$ to solve for y we obtain the generalized quantile function $y = y(p;\theta)$. This quantile function links change in the parameter with change in the variable y; the assumed model continuity provides the inverse. The local effect of the continuity at the observed data y^0 can then described by the $n \times p$ matrix of gradient vectors, called ancillary directions,

$$V = \left.\frac{\partial y(p;\theta)}{\partial \theta}\right|_{y^0,\hat{\theta}^0} \tag{.1}$$

which link change in the coordinates of θ, at $\hat{\theta}^0$, to change in the response, at y^0, for fixed p. A number of examples of the matrix V are given in Fraser et al. (2010, §3) and Brazzale et al. (2007, §8.4). The column vectors of V are tangent to the flow of probability under θ-change near $\hat{\theta}^0$; this flow defines the continuity-based ancillary contours concerning θ. Fraser et al. (2010) show that these vectors define a surface in the sample space that is ancillary to $O(n^{-1})$.

Our continuity assumption, which we view as intrinsic to a general approach to inference using approximate ancillarity, rules out unusual pivots as in the inverted Cauchy introduced in McCullagh (1992); see Fraser et al. (2010, Example 5). A different, although related, approach is needed for discrete responses y; see Davison et al. (2006) for discussion.

For vector parameters the approach builds on the presence of the quantile function presentation of the model and with independent vector coordinates may leave arbitrariness that can be addressed in other ways.

Given this matrix of ancillary directions V, a tangent exponential model with canonical parameter

$$\varphi(\theta) = \ell_{;V}(\theta; y^0),$$

can be constructed, where

$$\ell_{;V}(\theta; y^0) = \left.\frac{\partial \ell(\theta; y)}{\partial V}\right|_{\hat\theta^0, y^0}$$

is shorthand for the set of directional derivatives of $\ell(\theta; y)$ in the sample space, each direction determined by a row of V. The tangent exponential model is

$$f_{TEM}(s; \theta) = \exp\{\varphi(\theta)^T s + \ell(\theta; y^0)\} h(s), \tag{.2}$$

where $s \in \mathbb{R}^p$ has the same dimension as the parameter φ, and can be thought of as the score variable. The tangent exponential model was introduced in Fraser (1990); see also Reid and Fraser (2010, §2) and the references therein. The model was introduced mainly as a tool to obtain an r^* approximation for inference about a scalar component parameter, without the need to explicitly compute an ancillary density. Here we are using the tangent model as a descriptive device for obtaining a conditional density for inference about a scalar or vector parameter of interest, via saddlepoint approximations.

Example .1. Suppose y_i are independent observations from the curved exponential family $N(\psi; c^2\psi^2)$, where c is fixed. The ith component of the quantile vector p is $(y_i - \psi)/(c\psi)$, and the ith entry of the $n \times 1$ vector V is $y_i^0/\hat\psi^0$. Using this to define $\varphi(\theta)$ we have

$$\varphi(\theta) = \ell_{;V}(\theta; y^0) = \sum_{i=1}^n \frac{\partial \ell(\theta; y^0)}{\partial y_i} V_i = \frac{1}{c^2\psi}\sum(y_i^0/\hat\psi^0) - \frac{1}{c^2\psi^2}\sum\{(y_i^0)^2/\hat\psi^0\},$$

a linear combination of $1/\psi$ and $1/\psi^2$. In terms of the sufficient statistic $(\sum y_i, \sum y_i^2)$ an exact ancillary is $\sum y_i^2/(\sum y_i)^2$. The ancillary based on the V_i is consistent with this as both $\{y_i^0\}$ and $c\{y_i^0\}$ on the linear space $\mathcal{L}V$ give the same value to $\sum y_i^2/(\sum y_i)^2$.

Appendix B

(i) From likelihood to density by Taylor expansion. Example 4.1 is motivated by the usual Taylor series expansion of the log-likelihood function for a regular p-dimensional statistical model: the leading term is the log-likelihood for a normal distribution, with higher order terms that drop off as $n^{1/2}; n^{-1}; n^{-3/2}$; see for example, DiCiccio et al. (1990) and Cakmak et al. (1998). To simplify the calculations we introduce just one third derivative term: $a\lambda^2\chi/(2n^{1/2})$, where $a = \partial^3\ell/\partial\lambda^2\partial\chi$, evaluated at the expansion point. The resulting likelihood function can be inverted to provide an expression for the latent density $h(s,t)$, to $O(n^{-3/2})$, verifying 4.7:

$$\begin{aligned} g(s,t;\chi,\lambda) =& \frac{1}{2\pi}\exp\{-(s-\chi)^2/2 - (t-\lambda)^2/2 - a\chi\lambda^2/2n^{1/2}\} h(s,t) \\ =& \phi(s-\chi)\phi(t-\lambda)\{1 - a\chi\lambda^2/2n^{1/2} + a^2\chi^2\lambda^4/8n\} h(s,t) + O(n^{-3/2}) \\ =& \phi(s-\chi)\phi(t-\lambda)\{1 - a\chi\lambda^2/2n^{1/2} + a^2\chi^2\lambda^4/8n\} \\ & \times \{1 + as(t^2-1)/2n^{1/2} + a^2(s^2-1)(t^4-6t^2+3)/8n\} + O(n^{-3/2}); \end{aligned} \tag{.3}$$

the second equality uses $\exp(c/n^{1/2}) = 1 + c/n^{1/2} + c^2/2n + O(n^{-3/2})$, and the third equality uses $(1 - c/2n^{1/2} + c^2/8n)^{-1} = 1 + c/2n^{1/2} + c^2/8n + O(n^{-3/2})$ together with $E(x^2 - 1) = \theta^2, E(x^4 - 6x^2 + 3) = \theta^4$ when x follows a $N(\theta, 1)$ distribution.

(ii) From density to likelihood by Taylor expansion. The conditional model for s given t is available as the t-section of the density 4.6 and gives 4.8, up to a normalizing constant as:

$$\begin{aligned} g(s|t;\chi) =& c\phi(s-\chi)\{1 + as(t^2-1)/2n^{1/2} + a^2(s^2-1)(t^4-6t^2+3)/8n\} \\ =& \phi(s-\chi)\{1 + as(t^2-1)/2n^{1/2} + a^2(s^2-1)(t^4-6t^2+3)/8n\} \\ & \{1 + a\chi(t^2-1)/2n^{1/2} + a^2\chi^2(t^4-6t^2+3)/8n\}^{-1} \\ =& \phi(s-\chi)\{1 + as(t^2-1)/2n^{1/2} + a^2(s^2-1)(t^4-6t^2+3)/8n\} \\ & \{1 - a\chi(t^2-1)/2n^{1/2} + a^2\chi^2(t^4+2t^2-1)/8n\} \\ =& \phi(s-\chi)\{1 + as(t^2-1)/2n^{1/2} + a^2(s^2-1)(t^4-6t^2+3)/8n\} \\ & \exp\{-a\chi(t^2-1)/2n^{1/2} + a^2\chi^2(4t^2-2)/8n\} \end{aligned} \tag{.4}$$

The second equality comes by evaluating the constant c as the reciprocal of an integral with respect to s and uses $E(x) = \theta$ and $E(x^2-1) = \theta^2$ when x follows a $N(\theta, 1)$ distribution; the third equality comes from calculating the reciprocal of the factor coming from the preceding integration; and the fourth comes by taking the preceding to the exponent.

References

[1] Barndorff-Nielsen, O. E. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557-563.

[2] Barndorff-Nielsen, O. E. and D. R. Cox (1979). Edgeworth and saddlepoint approximations with statistical applications. *J. R. Statist. Soc. B* **41**, 187-220.

[3] Brazzale, A. R., A. C. Davison, and N. Reid (2007). *Applied Asymptotics. Cambridge: Cambridge University Press*.

[4] Cakmak, S., D. A. S. Fraser, P. McDunnough, N. Reid, and X. Yuan (1998). Likelihood centered asymptotic model: exponential and location model versions. *J. Statist. Planning and Inference* **66**, 211-222.

[5] Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals Math. Statist*. **46**, 21-31.

[6] Davison, A. C. (1988). Approximate conditional inference in generalized linear models. *J. R. Statist. Soc., B* **50**, 445-461.

[7] Davison, A. C., D. A. S. Fraser, and N. Reid (2006). Improved likelihood inference for discrete data. *J. R. Statist. Soc., B* **68**, 495-508.

[8] Davison, A. C., D. A. S. Fraser, N. Reid, and N. Sartori (2014). Accurate directional inference for vector parameters in linear exponential models. *J. Amer. Statist. Assoc.* **109**, 302–314.

[9] DiCiccio, T. J., C. A. Field, and D. A. S. Fraser (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika* **77**, 77-95.

[10] DiCiccio, T. J. and G. A. Young (2008). Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families. *Biometrika* **95**, 497-504.

[11] Fraser, A. M., D. A. S. Fraser, and A. M. Staicu (2010). Second order ancillary: A differential view with continuity. *Bernoulli* **16**, 1208-1223.

[12] Fraser, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77**, 65-76.

[13] Fraser, D. A. S. and H. Massam (1985). Conical tests: Observed levels of significance and confidence regions. *Statist. Hefte* **26**, 1-17.

[14] Fraser, D. A. S. and N. Reid (1993). Third order asymptotic models: Likelihood functions leading to accurate approximations for distribution functions. *Statist. Sinica* **3**, 67-82.

[15] Fraser, D. A. S. and N. Reid (1995). Ancillaries and third order significance. *Utilitas Mathematica* **47**, 33-53.

[16] Fraser, D. A. S. and N. Reid (2006). Assessing a vector parameter. *Student* **5**, 247-256.

[17] Fraser, D. A. S., N. Reid, and N. Sartori (2016). Accurate directional inference for vector parameters. *Submitted*.

[18] Fraser, D. A. S., N. Reid, and J. Wu (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249-264.

[19] Fraser, D. A. S. and J. Rousseau (2008). Studentization and deriving accurate p-values. *Biometrika* **95**, 1-16.

[20] Jensen, J. (1992). The modified signed likelihood statistic and saddlepoint approximations. *Biometrika* **79**, 693-703.

[21] Lee, S. and G. Young (2005). Parametric bootstrapping with nuisance parameters. *Statist. Prob. Lett.* **71**, 143-153.

[22] McCullagh, P. (1992). Conditional inference and Cauchy models. *Biometrika* **79**, 247-259.

[23] Reid, N. and D. A. S. Fraser (2010). Mean likelihood and higher order approximations. *Biometrika* **97**, 159-170.

[24] Skovgaard, I. (1988). Saddlepoint expansions for directional test probabilities. *J.R. Statist. Soc. B* **50**, 269-280.