Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

LIKELIHOOD FROM ESTIMATING FUNCTIONS

P. A. Mykland
University of Chicago

ABSTRACT

Likelihoods based on estimating equations have typically been designed to retain some, but not all, of the desirable properties of true likelihood. We discuss how these differences in intention affect a number of commonly used procedures, and, in particular, how various likelihoods live up to the ideal: true likelihood.

Key Words: Accurary, adaptive estimation, conditionality, dual likelihood, efficiency, empirical likelihood, projective likelihood, quasi-likelihood.

1 Introduction

Likelihood inference has the nice property that it solves several different problems at the same time. Such inference is first order efficient. It has nice accuracy properties, in particular when using the likelihood ratio statistic, its signed square root R, and the associated R^* (Barndorff-Nielsen (1986)) statistic. The accuracy is both unconditional and conditional (McCullagh (1984), Jensen (1992, 1997)). One gets a likelihood surface. And likelihood incorporates notions of inferential correctness from both the frequentist and Bayesian viewpoints.

This concatenation of desirable features would seem hard to replicate in a less than full parametric setting. Most approaches attempt to solve one of the above problems rather than all of them at the same time. Quasi- and projective likelihood (Godambe (1960), Wedderburn (1974), Godambe and Heyde (1987), McLeish and Small (1992)) and adaptive inference (going back to Beran (1974), Sacks (1975) and Stone (1975); see Bickel, Klaassen, Ritov and Wellner (1993) for a quite comprehensive account) are, essentially, solutions to the efficiency problem. Dual likelihood and empirical likelihood for estimating equations (Kolaczyk (1994), Qin and Lawless (1994), Mykland (1995)) are solutions to the unconditional accuracy problem. The original empirical likelihood (Owen (1988, 1990), see also DiCiccio and Romano (1989) and DiCiccio, Hall and Romano (1991)) appears to have been motivated by both considerations.

To what exent do these methods cope with the problems which they were *not* designed to solve?

2 Unintended Properties

As far as the author is aware, this question has not been heavily studied. In the following, we summarize what appears to be known about the properties mentioned above.

(i) **Optimality**

This is the high ground of quasi-likelihood and adaptive inference. If an empirical or dual likelihood is based on a quasi-score, it will have the same asymptotic efficiency as the score itself (Kolaczyk (1994), Mykland (1995)). One can also base such likelihoods on other scores, but this would be unnatural if one knows what the second moment structure is like. We are not aware of any work concerning any possible connection between adaptive inference and empirical or dual likelihood.

(ii) **Unconditional Accuracy**

This is what empirical and dual likelihood are good at, though things start breaking down in the presence of nuisance parameters (Lazar (1996), Lazar and Mykland (1996), Mykland (1996)). The quasi-log likelihood does not typically satisfy Bartlett identities of order higher than 2, so the accuracy properties of the R statistic and its cousins do not hold (cf. Mykland (1996)). If one wishes these properties, one can instead use projective likelihood (McLeish and Small (1992)), which is based on the same inferential ideas as quasi-likelihood. By virtue of the projective likelihood being a true Radon-Nikodym derivative, accuracy will be as for likelihood. There is no free lunch, however, as we shall see next. Little is known about the accuracy of adaptive inference beyond first order.

(iii) **Likelihood Surface**

This exists for quasi- and empirical likelihood, and can presumably also be defined in the context of adaptive inference. For the other two approaches which we are discussing, the issue is more problematic. Projective likelihood needs a reference parameter value; the log is on the form $l_{\theta_0}(\theta)$, and, typically, $l_{\theta_0}(\theta_1)+l_{\theta_1}(\theta_2)$ is not the same as $l_{\theta_0}(\theta_2)$. One can locate the reference value at the MPLE, but this is not a completely satisfying solution.

For dual likelihood, this question is not fully explored. If it is a dual criterion function to a nonparametric likelihood (empirical or point process, cf. Section 6 in Mykland (1995)), one can presumably use the surface from the nonparametric quantity. For a 'pure' dual likelihood (based on an estimating equation only, with no nonparametric counterpart, such as Aalen's (1980, 1989) linear regression), we do not know whether a likelihood surface exists.

(iv) **Conditional Properties**

We are not aware of any work in this direction, so here is a first stab at this. Let us suppose that the quasi-score and its derivative is correctly specified, in the sense that they coincide with the first two derivatives in the 'true' (unknown) log likelihood. In this case, the quasi and projective R statistics are, obviously, second order locally sufficient in the sense of McCullagh (1984). This, however, is not the case for empirical or dual R.

In regular cases, the argument goes as follows: by the Hájek-LeCam convolution theorem (see, *e.g.*, Hájek (1969)), $\bar{A} = [\dot{l}, \dot{l}](\hat{\theta}) + \ddot{l}(\hat{\theta})$ is first order ancillary. In view of McCullagh (1984), there is a B, $B = O_p(1)$, so that $A = \bar{A} + B$ is second order ancillary. Hence, by McCullagh (1984), $\operatorname{cov}(A, \ddot{l}(\theta)) + \operatorname{cum}(A, \dot{l}(\theta), \dot{l}(\theta)) = o(n)$, and so

$$\operatorname{cov}(A, -[\dot{l}, \dot{l}](\theta)) + \operatorname{cum}(A, \dot{l}(\theta), \dot{l}(\theta)) = -\operatorname{var}(\bar{A}) + o(n).$$

This expression is $o(n)$ only if $\operatorname{var}(\bar{A})$ is $o(n)$, which, under standard assumptions, translates into $\bar{A} = O_p(1)$. Since $\bar{A} = [\dot{l}, \dot{l}](\theta) + \ddot{l}(\theta) + c\dot{l}(\theta) + O_p(1)$, this means in turn that

$$l_d(\theta + \delta) - l_d(\theta) = l_d(\theta + \mu) - l_d(\theta) + O_p(\mu^3),$$

where l_d and l are the dual and true likelihood, respectively, and where $\mu = \delta - \delta^2/2$. Hence, by McCullagh (1984), the dual (and hence the empirical) R statistic is second order locally sufficient only if the dual likelihood coincides with the true one (and hence with the quasi-likelihood) to second order locally at θ.

This leads, *inter alia*, to the conclusion that the dual R is unconditionally more accurate than the quasi-R, but conditionally less so!

What if the second order structure is not known? One may then have a choice between overdispersed quasi- and empirical/dual likelihood. In this case, things are less clear. The issue is pursued in Lazar (1996).

There are a number of other questions here. To mention a few: What about adaptive estimation? And the above only tackles second order local sufficiency. What about the large deviation properties documented in Skovgaard (1990, 1996), Jensen (1992, 1997) and Barndorff-Nielsen and Wood (1995)?

3 Conclusion

It has hopefully been illustrated in the above that there are a substantial number of unresolved issues in this area. Even more fundamentally, there are also more questions which need to be asked. If likelihood is the gold standard, then what are the properties of likelihood anyway? New ones keep being discovered, as the rich recent literature on the subject can testify. And are there criterion functions yet to be discovered which come closer to the gold standard than the ones we have discussed?

Acknowledgements

This research was supported in part by National Science Foundation grants DMS 93-05601 and DMS 96-26266, and Army Research Office grant DAAH04-95-1-0105. The manuscript was prepared using computer facilities supported in part by the National Science Foundation grants DMS 89-05292, and DMS 87-03942 awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

References

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes, *Lect. N. Statist.* **2** 1–25 (Springer, New York).

Aalen, O. (1989). A linear regression model for the analysis of life times, *Statist. in Medicine* **8** 907–925.

Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio, to appear in *Biometrika* **73** 307–322.

Barndorff-Nielsen, O. E., and Wood, A. T. A. (to appear). On large deviations and choice of ancillary for p^* and the modified directed likelihood. *Bernoulli.*

Beran, R. (1974). Asymptotically efficient adaptive rank estimates in location models, *Ann. Statist.* **2** 63–74.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models* (Johns Hopkins).

DiCiccio, T.J., Hall, P., and Romano, J.P. (1991). Empirical likelihood is Bartlett-correctable, *Ann. Statist.* **19** 1053–1061.

DiCiccio, T.J., and Romano, J.P. (1989). On adjustments based on the signed root of the empirical likelihood ratio statistic, *Biometrika* **76** 447–456.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation, *Ann. Math. Statist.* **31** 1208–1212.

Godambe, V.P., and Heyde, C.C. (1987). Quasi-likelihood and optimal estimation, *Int. Statist. Rev.* **55** 231–244.

Hájek, J. (1969). A characterization of limiting distributions of regular estimates, *Z. Wahrsch. verw. Gebiete* **14** 323–330.

Jensen, J. L. (1992). The modified signed likelihood statistic and saddlepoint approximations. *Biometrika* **79** 693–703.

Jensen, J. L. (1997). A simple derivation of r^* for cuved exponential families. *Scand. J. Statist.* **24** 33–46.

Kolaczyk, E.D. (1994). Empirical likelihood for generalized linear models, *Statistica Sinica* **4** 199–218.

Lazar, N. (1996). *Some Inferential Aspects of Empirical Likelihood.* Ph.D. dissertation, University of Chicago.

Lazar, N., and Mykland, P. A. (1996). Empirical likelihood in the presence of nuisance parameters, Technical report no. 400, Dept. of Statistics, University of Chicago.

McCullagh, P. (1984). Local sufficiency. *Biometrika* **71** 233–244.

McLeish, D.L., and Small, C.G. (1992). A projected likelihood function for semiparametric models, *Biometrika* **79** 93–102.

Mykland, P.A. (1995). Dual likelihood. *Ann. Statist.* **23** 396–421.

Mykland, P.A. (1996). The accuracy of likelihood, Technical report no. 420, Department of Statistics, University of Chicago.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional, *Biometrika* **75** 237–249.

Owen, A.B. (1990). Empirical likelihood ratio confidence regions, *Ann. Statist.* **18** 90–120.

Qin, J., and Lawless, J. (1994). Empirical likelihood and general estimating equations, *Ann. Statist.* **22** 300–325.

Sacks, J. (1975). An asymptotically efficient sequence of estimators of a location parameter, *Ann. Statist.* **3** 285-298.

Skovgaard, Ib (1990). On the density of minimum contrast estimators. *Ann. Statist.* **18** 779–789.

Skovgaard, Ib M. (1996). An explicit large-deviation approximation to one-parameter tests. Bernoulli **2** 145-165.

Stone, C.J. (1975). Adaptive maximum likelihood estimators of a location parameter, *Ann. Statist.* **3** 267–284.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61** 439–447.