Geometry, Likelihood and Inference: the Work of Bruce G. Lindsay

Nancy Reid*

Abstract

Bruce Lindsay's approach to many problems in inference was very geometric. This gave him great insight into the structure of models, although it was sometimes hard for the rest of us to keep up. In this talk I will consider a few of his many contributions to the theory of likelihood inference, with a view to highlighting how his insight informed work that followed.

Key Words: mixture models, composite likelihood, memorial, projection, conditional inference

1. Introduction

Bruce Lindsay and I started our careers just one year apart, unbeknownst to me at the time. I met him quite serendipitously; having arrived as a new Assistant Professor at the University of British Columbia in 1980, I quickly realized that the faculty there created a lively collaborative environment by inviting visitors to the department during the summer months. Soon after that I came across Bruce's paper (Lindsay, 1982), and thought "there's someone I could talk to!", and proceeded to prepare a 'cold-call' invitation. It turned out, as these things often do, that my choice of paper, and author, was less accidental than I thought. Bruce had spent the year 1978-1979 at Imperial College, London, where I spent 1979-1980 as a postdoctoral fellow, so we were clearly moulded by the same environment.

Even luckier, Bruce had been planning in any case to be on the West Coast in the summer of 1984, and had signed up, with a college friend, for a guided trip to the summit of Mount Rainier in Washington State. He graciously asked me if I would be interested in joining them for this trip, and I spent the weeks between accepting, and the climb itself, trying to get fit enough to manage it. Certainly for me, and Laura tells me for Bruce as well, this was a very memorable experience, and loomed large in our personal histories.

While we didn't end up writing any papers together, we had close and overlapping interests in many aspects of statistical theory throughout our careers. As other speakers in the memorial session noted, Bruce had a low-key personal manner, and a stubbornly independent streak. As a result he was highly unimpressed with the fashion of the day in research, and had a succinct way of keeping me grounded. For example, when I won the Presidents' Award, and was very naturally feeling quite pleased with myself, he wrote simply "Congratulations. Don't let it go to your head".

I chose for the presentation at the Joint Statistical Meetings to describe briefly three of Bruce's papers. They were selected from a very long list of important contributions partly because they interested me, but also because they give some flavour of the range of his ideas, his continual emphasis on geometric thinking, and in my opinion some of the very interesting ideas in statistical theory that emerged during our careers. We were both very fortunate to be part of a remarkable period of advance and excitement in what used to be called "mathematical statistics"; the theory and foundations of statistical inference, and its importance for statistical practice.

*Department of Statistical Sciences, University of Toronto, 100 St. George St., Toronto Canada M5S 3G3

2. Conditional Score Functions

Lindsay (1982) is a key contribution to the development of conditional inference, which was being much discussed at the time. Conditional inference is important in the theory of statistics in ensuring that inferences are relevant to the data that has been observed, and is intimately connected to notions of both sufficiency and ancillarity. Bruce's work in this paper considers conditional inference from the point of view of estimating equations and score functions. He particularly focusses on what are now called "Neyman-Scott problems", which assumes the data are independent vectors $X_i \sim f(x; \theta, \phi_i), i = 1, \ldots, N$. Each additional vector observation brings in a new nuisance parameter, but the parameter of interest θ is common to all the data. He considers how to improve on the usual estimating function for θ:

$$\Sigma \widehat{U}_i(\theta) = 0,$$

where $\widehat{U}_i(\theta) = U_i\{\theta, \hat{\phi}(\theta)\}$, $U_i(\theta, \phi_i) = (\partial/\partial\theta) \log f(x_i; \theta, \phi_i)$ and $\hat{\phi}(\theta)$ is typically the constrained maximum likelihood estimate of $\phi = (\phi_1, \ldots, \phi_N)$. He suggests instead replacing the ith estimated score function by the *conditional score function*

$$W_i(\theta) \equiv U_i(\theta, \phi_i) - E_\theta(U_i \mid S_i),$$

where it is assumed that we have available a statistic S_i which is sufficient for ϕ_i, when θ is fixed, thus ensuring that W_i only depends on θ. Importantly he considers the case that S_i is a function only of the data, such as in inference for the canonical parameter in an exponential family model, and the case that S_i is a function of θ, which is more delicate. In this latter case $W_i(\theta) = \phi[S'\theta) - E\{S'(\theta) \mid S(\theta)\}]$; this arises for example in inference for the ratio of two canonical parameters of an exponential family model.

To study the structure of these conditional score functions Bruce goes on to consider some general properties of estimating functions, and these have become part of the standard literature of inference. An estimating function is used to estimate θ via the equation

$$H(\theta; y) = 0,$$

where y represents the totality of the observed data; in the notation above, the set of N vectors x_i. This estimating function is unbiased if $E\{H(\theta; Y)\} = 0$, and is *information unbiased* if

$$E\{H'(\theta) + H^2(\theta)\} = 0.$$

The *information in* H is defined as

$$i_H = \{E(H')\}^2 / \{E(H^2)\}.$$

All expectations are taken under the model $f(\cdot; \theta, \phi)$. Bruce then shows that the conditional score function $\Sigma W_i(\theta)$ is information unbiased, at the true value of the parameter (θ_0, ϕ_0), and develops conditions under which, among all unbiased estimating functions, $\Sigma W_i(\theta)$ maximizes i_H, relating this to the score function from the partial likelihood of Cox (1975).

Bruce then turns to the challenge of dealing with the nuisance parameter, as W will depend on this if the conditioning statistic S is a function of the parameter of interest θ. Replacing ϕ by an estimate $\tilde{\phi}$, and writing

$$\widetilde{W}(\theta) = \Sigma W_i(\theta, \widetilde{\phi}),$$

he notes that optimality, in the sense of minimizing i_H, cannot be attained in most settings. He defines a class of such estimated estimating equations by imposing a form of conditional information unbiasedness, finds the optimal solution in this restricted class, and notes:

> Since the solution depends on parameter c, an entire class of estimated conditional scores exists. In the following examples elementary considerations ... and *statistical good sense* [my emphasis] seem to point to a single value of c.

You can see the influence of the "Imperial College school of statistics" in this phrase!

The paper concludes with short sections that have a number of intriguing suggestions and results. It is noted that, for example,

$$i_F(\theta) = i_W(\theta) + E\{(...)^2\},$$

where $i_F(\theta)$ is the Fisher information, known by results of Godambe (196x) to be the maximal information in an estimating function, and $i_W(\theta)$ is the information in the conditional score function. About the positive term added to that Bruce writes that it is "inherently plausible" that it would converge to 0.

He also suggests applying the results of the paper to the model

$$L(\theta, \tau) = \prod_{i=1}^{N} \int f(x_i; \theta, \phi) dQ_\tau(\phi),$$

which is the mixture model that formed such a large part of Bruce's work, and for which he obtained quite remarkable and detailed inferential results.

Lindsay (1982) has become part of the canon in the literature on conditional and marginal inference, and Bruce followed up on this in several different directions, including Waterman and Lindsay (1996a, 1996b, 1999), Li et al. (2003), Lindsay and Qu (2003), Hui and Lindsay (2010).

3. Estimating the Number of Classes

As with the paper in the previous section, I had several personal reasons for selecting Mao and Lindsay (2004): first, it was published in the journal of the Statistical Society of Canada. Second, in Joe and Reid (1985) we addressed a simpler, but related, problem of estimating the number of faults in a reliability system, and there was a lot of interest throughout the 1980s in the estimation of the "sizes" of things: the number of species in a region (Efron and Thisted, 1976), the size of an ecological population (Raftery, 1988) and so on. The inference problem is difficult because the usual regularity conditions do not apply: the parameter space is discrete, and the log-likelihood function is often monotone increasing or decreasing. Even when the likelihood function is unimodal, the confidence intervals obtained using the usual inferential theory tend to be either much too narrow or much too wide. The fact that Bruce found new insights into this class of problems as late as 2004 illustrates his originality, and his interest in pursuing problems that are, often only temporarily, 'out of fashion'. Estimating the number of classes is again appearing in important applications to clustering.

The formulation in this paper is fairly general, considering that there are m populations of individuals, and within all these m populations there are an unknown number c of classes. Letting $i = 1, \ldots, c$ index classes, and $k = 1, \ldots, m$ index populations, Y_{ik} denotes the number of individuals in population k observed to be in class i. The particular application mentioned in this paper arises in ecology, where the classes are species, and the populations are actually different methods of identifying species.

The frequency vector for the ith class is $Y_i = (Y_{i1}, \ldots, Y_{im})'$, and the complete sample is the matrix with rows Y_i, with the complication that rows of all 0s are not observed, so that the matrix actually has n rows, not c.

The model that is developed in this paper assumes

$$Y_{ik} \sim Poisson(\lambda_{ik}),$$

where the vector of rate parameters $\lambda_i = (\lambda_{i1}, \ldots, \lambda_{im})'$ ranges over $[0, \infty)^m$, with the constraint that the zero-vector is excluded. Dependence among the populations is captured by dependence in the components of λ_i, leading to the semi-parametric mixture model

$$Y \sim \prod_{i=1}^{c} g_P(y_i),$$

where

$$g_P(y_i) = \int \prod_{k=1}^{m} \frac{e^{-\lambda_{ik}} \lambda_{ik}^{y_{ik}}}{y_{ik}!} dP(\lambda_i).$$

and $\lambda_i \sim P(\cdot)$.

The likelihood function for this semi-parametric model can be factored as

$$L(c, P) = L_{marg}(c, P \mid n) \times L_{cond}\{P \mid (n_x)\},$$

where n_x is the set of the observed counts, excluding the rows of the Y matrix that are identically 0, n is the number of observed classes, and $c - n$ is the number of unobserved classes. The marginal likelihood has a simple Binomial form,

$$\binom{c}{n} \pi_P^{c-n} (1 - \pi_P)^n,$$

where the sample size c is the parameter of interest. The maximum likelihood estimate from this marginal likelihood is

$$\hat{c}_P = n(1 + \frac{\pi_P}{1 - \pi_P}),$$

depending only on the odds ratio $\theta(P) = \pi_P/(1 - \pi_P)$.

The more difficult part is the conditional log-likelihood, which can be expressed using a standard mixture model after reparametrization from mixture density P to a related mixture density Q, with the parameter of interest to be estimated now denoted $\theta(Q)$. The authors show that the parameters in the model are identifiable, but the functional is 'badly discontinuous' in the sense of Donoho (1988), which has the following implications for inference:

- all estimates have either very high bias or very low variance
- we can only obtain lower confidence bounds
- upper confidence limits have very high probability of being infinite

These cautionary remarks serve as a counterpoint to the unthinking use of likelihood inference in applications, on the grounds that it usually works. Mao and Lindsay (2007) develop the theory of this further.

A final reason I chose this paper is to highlight part of the acknowledgment: "Professor Lockhart's amazing input to improve the presentation is particularly noteworthy." Richard Lockhart was the editor of the journal at the time, and is indeed well-known as a careful and helpful editor, but I also like to think of this as a tongue-in-cheek reference to the fact that the first versions of Bruce's papers are not always the clearest.

4. Confidence Distribution Sampling

Kim and Lindsay (2011) is essentially an algorithmic paper, providing a means of sampling in counterpoint to, for example, MCMC sampling from a posterior. I chose it again for partly personal reasons: it appeared in *Statistica Sinica* in 2011, and in the same year Bruce, Cristiano Varin, and I co-edited a special issue of that journal on composite likelihood that was initiated by Kung-Yee after a very successful workshop at the University of Warwick.

There is renewed interest in confidence distributions recently, as part of an effort to find common ground between Bayesian and non-Bayesian inference, and there were two sessions at JSM 2016 on "BFF": Bayes-fiducial-frequentist.

In this paper the authors assume that an inference function $H(y;\theta)$ is available that provides a point estimate $\hat{\theta}$ and an estimate $V_{\hat{\theta}}$ of the covariance matrix of $\hat{\theta}$. Although the notation H was used in Lindsay (1982) as an arbitrary estimating function, thus generalizing the score function, here it is a generalization of the log-likelihood function. Thus the log-likelihood function itself is an inference function; other examples treated in the paper are Owens (1988) empirical likelihood function, the quadratic inference function (Lindsay and Qu, 2003; Qu et al., 2000), and the quadratic form given by Raos score test.

They assume that the inference function permits testing values of θ at arbitrary levels α, via sets of the form

$$C_H \equiv \{\theta : H(\theta; y) - H(\hat{\theta}; y) \leq c_\alpha\},$$

and take as a definition of an asymptotic confidence distribution a function for which

$$P_{CD}(\theta \in C_H \mid y) = 1 - \alpha.$$

The goal in the paper is to sample from this confidence distribution, in order to be able to visualize it. Visualization is currently becoming a prominent aspect of the world of big data.

The algorithm proposed has something of the flavour of indirect inference (cf. e.g. Jiang and Turnbull, 2004). Given $\hat{\theta}$ and $V_{\hat{\theta}}$, obtained from the inference function H:

- simulate $z \sim N_p(0, I)$ and compute $\alpha(z) = Pr(\chi^2_p \geq z'z)$
- solve for ϵ: $H(\hat{\theta} + \epsilon V^{1/2} z) - H(\hat{\theta}; y) = z'z$
- $\theta(z) = \hat{\theta} + \epsilon V^{1/2} z$ is in the confidence region, on a line towards the boundary
- repeat

Figure 1 in Kim and Lindsay (2011) illustrates the sampling and the visualization, and describes visually their more sophisticated version of the algorithm that pushes the points in the confidence region towards the boundary, i.e. ensuring that $\alpha(z)$ in step 2 of the algorithm is closer to the desired 5% level, for example, so that less time is spent sampling θ from the interior of the confidence set. Not surprisingly, one of the main illustrations of the technology in the paper is to a mixture model!

5. Conclusion

This is a highly abbreviated look at just a tiny portion of Bruce's many contributions to statistical theory and practice. I am glad to have been able to count myself among his friends, and I learned a great deal from reading his papers and talking to him about research. As one of the leaders of our generation of statisticians he is much missed. And of course, I

am forever grateful to him for creating the opportunity to spend the rest of my life reminding myself and others that I managed to get up Mt. Rainier, back in the day. Not that Bruce would be so easily impressed.

REFERENCES

Donoho, D. (1988), "One-sided Inference About Functionals of a Density", *The Annals of Statistics*, 16, 1390–1420.

Efron, B. and Thisted, R. (1976), "Estimating the Number of Unseen Species: How Many Words did Shakespeare Know?", *Biometrika*, 63, 435–447.

Jiang, W. and Turnbull, B. (2004), "The Indirect Method: Inference Based on Intermediate Statistics", *Statistical Science*, 19, 239–263.

Joe, H. and Reid, N. (1985), "Estimating the Number of Faults in a System", *Journal of the American Statistical Association*, 80, 222–226.

Kim, D. and Lindsay, B.G. (2011), "Using Confidence Distribution Sampling to Visualize Confidence Sets", *Statistical Sinica*, 21, 923–948.

Li, H., Lindsay, B.G. and Waterman, R. P. (2003), " Efficiency of Projected Score Methods in Rectangular Array Asymptotics", *Journal of the Royal Statistical Society, Series B*, 65, 191–208.

Lindsay, B.G. (1982), "Conditional Score Functions: Some Optimality Results," *Biometrika*, 69, 505–512.

Lindsay, B.G. and Qu, A. (2003), "Inference Functions and Quadratic Score Tests", *Statistical Science*, 18, 294–410.

Lindsay, B.G. and Waterman, R.P. (1999), "A Simple Measure of Second Order Information Loss due to the Presence of Nuisance Parameters", in *Asymptotics, Nonparametrics and Time Series: a tribute to Madan Lal Puri*, Ed. S. Ghosh, New York: Marcel Dekker, Inc.

Mao, C.-X. and Lindsay, B.G. (2004), "Estimating the Number of Classes in Multiple Populations: a Geometric Analysis," *The Canadian Journal of Statistics*, 32, 303–314.

Mao, C.-X. and Lindsay, B.G. (2007), "Estimating the Number of Classes, *Annals of Statistics*, 35, 917–930.

Owen, A. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional, *Biometrika*, 75, 237–249.

Qu, A., Lindsay, B.G. and Li, B. (2000), "Improving Generalised Estimating Equations Using Quadratic Inference Functions", *Biometrika*, 87, 823–836.

Raftery, A.E. (1988), "Inference for the Binomial N Parameter: A Hierarchical Bayes Approach", *Biometrika*, 75, 223–228.

Waterman, R.P. and Lindsay, B.G. (1996a), "A Simple and Accurate Method for Approximate Conditional Inference in Exponential Family Problems," *Journal of the Royal Statistical Society, Series B*, 58, 177–189.

Waterman, R.P. and Lindsay, B.G. (1996b), "Projective Score Methods in Nuisance Parameter Problems", *Biometrika*, 83, 1–15.