On the Expected Amount of Information from a Non-Linear Model

Nicholas G. Polson

Journal of the Royal Statistical Society. Series B (Methodological), Volume 54, Issue 3 (1992), 889-895.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281992%2954%3A3%3C889%3AOTEAOI%3E2.0.CO%3B2-W

J. R. Statist. Soc. B (1992)
54, *No.* 3, *pp.* 889–895

On the Expected Amount of Information from a Non-linear Model

By NICHOLAS G. POLSON†

University of Chicago, USA

[Received March 1990. Final revision August 1991]

SUMMARY

An approach to quantifying the amount of information that an experimenter expects to learn from a non-linear model is given. An expected utility for an experiment, $\mathscr{E}$, motivated by the asymptotic form of Shannon information gain between prior and posterior, is defined. This leads to a characterization of the experimenter who expects to learn the most from a non-linear model. Such an experimenter has the design-dependent Jeffreys prior. Sufficient regularity conditions for the equivalence with asymptotic Shannon information gain are given. An application to the optimal selection of sample size from a model with exponential family errors and the Michaelis–Menten model is discussed. A link between the regularity conditions for asymptotic posterior normality and the Jeffreys prior is given.

Keywords: BAYESIAN INFERENCE; DESIGN; JEFFREYS PRIOR; NON-LINEAR REGRESSION; REFERENCE PRIOR; SAMPLE SIZE; SHANNON INFORMATION

1. INTRODUCTION

Suppose that an experimenter can learn about a parameter of interest θ from an experiment $\mathscr{E}$ defined by observing data $\mathbf{y} = (y_1, \ldots, y_n)$ from the model $y_i = \eta(x_i, \theta) + \epsilon_i$. Here $\eta(\ ,\ )$ is a given response surface, x_i are design points and ϵ_i, $1 \leqslant i \leqslant n$, are independent and identically distributed with density $f(\ )$. Let X denote the set of design points, $\eta_i = \eta(x_i, \theta)$ and let $I_f = \int f'^2/f \, \mathrm{d}\nu$ be Fisher information where ν is a dominating measure. The density f need not be normal. The objective of experimentation is to learn about θ, which has prior density $p(\theta) > 0$, $\theta \in \Theta \subset \mathbf{R}^k$.

The purpose of this paper is to quantify the amount of information that an experimenter can expect to learn about the parameter θ through experimentation. Section 2 proposes an expected utility, $U^{\Theta}(\mathscr{E})$, quantifying the expected amount of information. The main application of $U^{\Theta}(\mathscr{E})$ is to the comparison of experiments (DeGroot, 1962), as it induces an order on experiments, in that $\mathscr{E}_1 \leq \mathscr{E}_2$, if $U^{\Theta}(\mathscr{E}_1) \leqslant U^{\Theta}(\mathscr{E}_2)$. In particular, it applies to spaces of design measures, response surfaces and the choice of sample size. Only the last will be considered here. The application to non-linear design problems (Chaloner and Larntz, 1989) and model choice problems are discussed in Polson (1988). The motivation for the expected utility $U^{\Theta}(\mathscr{E})$ is given by a suitably normalized form of Shannon information gain (Polson, 1988). The Shannon information gain was first proposed in a statistical setting as a measure of information of an experiment by Lindley (1956) and has been applied in many contexts, e.g. the design of linear models (Stone, 1959; Smith and Verdinelli, 1980) and optimal allocation problems (Brooks, 1980, 1987). For independent and identically distributed observations Bernardo (1979) uses asymptotic Shannon information gain for interpreting the Jeffreys prior as a reference prior. Here the

†*Address for correspondence*: Graduate School of Business, University of Chicago, Chicago, IL 60637, USA.

 0035–9246/92/54889 $2.00

observations are not identically distributed and Appendix A gives sufficient regularity conditions, based on the likelihood ratio process, for determining the asymptotic properties of Shannon information gain (Ibragimov and H'asminsky, 1973, 1975a, b) which avoids the heuristic approximation based on asymptotic posterior normality used by Bernardo (1979). It is shown that the experimenter who expects to learn the most, under $U^{\Theta}(\mathscr{E})$, adopts a Jeffreys prior. Care must be exercised when adopting this automatic rule as it depends on both the design and the sample size.

Section 3 considers two applications: first, the error density is assumed to belong to the exponential power family and, secondly, the choice of sample size to achieve a fixed level of $U^{\Theta}(\mathscr{E})$ for the Michaelis–Menten model with t errors is determined. Such a criterion is closely related to the criterion proposed by Lindley (1956) of sampling to attain a fixed level of the Shannon information gain.

Let $f_i(y_i|\theta) = f(y_i - \eta(x_i, \theta))$ be the likelihood of observation y_i and denote the full likelihood by

$$f(\mathbf{y}|\theta) = \prod_{i=1}^{n} f_i(y_i|\theta).$$

Suppose that each function f_i is absolutely continuous in θ and that the vector of derivatives $f_i'(y|\theta) = \partial f_i/\partial\theta$ exists for almost all y with respect to the measure ν and the integrals

$$I_i(\theta) = \int_{\{y: f_i(y|\theta) \neq 0\}} \frac{f_i'(y|\theta)^2}{f_i(y|\theta)} \nu(\mathrm{d}y)$$

are defined and finite $\forall\, \theta \in \Theta$ where the prime denotes the derivative with respect to θ. Hence,

$$I_i(\theta) = I_f \frac{\partial \eta_i^{\mathrm{T}}}{\partial\theta} \frac{\partial \eta_i}{\partial\theta}. \tag{1}$$

Let $I_X(\theta) = \Sigma_{i=1}^{n} I_i(\theta)$ denote Fisher information.

2. EXPECTED UTILITY

The Shannon information gain (Lindley, 1956) from an experiment $\mathscr{E}$, denoted by $I^{\Theta}(\mathscr{E})$, is defined by

$$I^{\Theta}(\mathscr{E}) = E_{\mathbf{y},\theta}\left[\log\left\{\frac{p(\theta|X, \mathbf{y})}{p(\theta)}\right\}\right] \tag{2}$$

where E_Y denotes the expectation with respect to the random variable Y and $p(\theta|X, \mathbf{y})$ is the posterior obtained after observing n observations under design X. A probabilistic tool for exploring the asymptotics of equation (2) is the likelihood ratio process $Z_{n,\theta}(\alpha)$, $\alpha \in \mathbf{R}^k$, which is defined by

$$Z_{n,\theta}(\alpha) = \prod_{i=1}^{n} \frac{f_i(y_i|\theta + \psi_{X,n}(\theta)\alpha)}{f_i(y_i|\theta)} \tag{3}$$

and the normalizing sequence $\psi_{X,n}(\theta)$ is given by $\psi_{X,n}^{-2}(\theta) = I_X(\theta)$. See Basawa and Prakasa Rao (1980), Ibragimov and H'asminsky (1975a, b) and Le Cam (1986).

Appendix A considers the asymptotics of equation (2) in this context of non-identically distributed observations based on the independent and identically distributed case considered in Ibragimov and H'asminsky (1973). The asymptotic result is

$$\lim_{n\to\infty}\left\{I^{\Theta}(\mathscr{S})-E_{\theta}\left(\log\left|\sum_{i=1}^{n}\frac{\partial\eta_i^{\mathrm{T}}}{\partial\theta}\frac{\partial\eta_i}{\partial\theta}\right|^{1/2}\right)\right\}=H\{p(\theta)\}+\frac{k}{2}\log\left(\frac{I_f}{2\pi e}\right), \tag{4}$$

where the second term is related to the expected logarithm of the determinant of Fisher information and $H\{p(\theta)\} = -\int p(\theta)\log p(\theta)\,\mathrm{d}\theta$ denotes the entropy functional. A heuristic interpretation of asymptotic Shannon information, quantified by equation (4), is the following: for a consistent model the posterior converges to a point mass at θ and so measuring the asymptotic gain in information from the prior to the posterior can be interpreted as the missing information about θ under the given experiment $\mathscr{S}$ and prior $p(\theta)$. See Bernardo (1979) for further discussion.

Equation (4) provides the basis for the following definition.

Definition 1. For each $\mathscr{S}$, define $U^{\Theta}:\mathscr{S}\mapsto U^{\Theta}(\mathscr{S})$ by

$$U^{\Theta}(\mathscr{S})=E_{\theta}\left(\log\left|\sum_{i=1}^{n}\frac{\partial\eta_i^{\mathrm{T}}}{\partial\theta}\frac{\partial\eta_i}{\partial\theta}\right|^{1/2}\right)+H\{p(\theta)\}+\frac{k}{2}\log\left(\frac{I_f}{2\pi e}\right). \tag{5}$$

The following lemma characterizes the design-dependent Jeffreys prior, denoted by $\pi_{X,n}(\theta)$, by establishing an upper bound for the expected utility defined in equation (5). Bernardo (1979) characterizes the Jeffreys prior for independent and identically distributed observations.

Lemma 1. Suppose that $|I_X(\theta)|^{1/2}\in L^1(\Theta)$ and define

$$\pi_{X,n}(\theta) = |I_X(\theta)|^{1/2}/\int|I_X(\theta)|^{1/2}\,\mathrm{d}\theta.$$

Then, the following inequality holds:

$$U^{\Theta}(\mathscr{S})\leqslant\log\left\{\int|I_X(\theta)|^{1/2}\,\mathrm{d}\theta\right\}+\frac{k}{2}\log\left(\frac{I_f}{2\pi e}\right)$$

with equality if and only if $p(\theta)\propto|I_X(\theta)|^{1/2}$.

Proof. From equation (1), summing over i, and taking the determinant and square root, we see that

$$\pi_{X,n}(\theta)\propto\left|\sum_{i=1}^{n}\frac{\partial\eta_i^{\mathrm{T}}}{\partial\theta}\frac{\partial\eta_i}{\partial\theta}\right|^{1/2}.$$

Now, normalizing and substituting into equation (5), we obtain

$$U^{\Theta}(\mathscr{S}) = E_{\theta}\left[\log\left\{\frac{\pi_{X,n}(\theta)}{p(\theta)}\right\}\right]+\log\left\{\int|I_X(\theta)|^{1/2}\,\mathrm{d}\theta\right\}+\frac{k}{2}\log\left(\frac{I_f}{2\pi e}\right).$$

The first term on the right-hand side is minus a Kullback–Leibler distance between the density $\pi_{X,n}(\theta)$ and the prior $p(\theta)$. This is less than or equal to 0, with equality if and only if $p(\theta) = \pi_{X,n}(\theta)$ (Kullback, 1959). Hence, we obtain the desired result.

Therefore, the experimenter with a Jeffreys prior expects to learn the most from experimentation, i.e. attains $\max_{p(\ )}\{U^{\Theta}(\mathscr{E})\}$.

For smooth models in the independent and identically distributed case, the Fisher information is linear in n and the experimenter learns at a rate $\frac{1}{2}\log n$. This is not necessarily so for a non-linear model; the rate is design dependent and is given by $E_\theta\{\log|\psi_{X,n}(\theta)|\}$. For example, suppose that $\eta(x_j,\theta) = \sin(j\theta)$, $\epsilon_i \sim N(0,1)$. Then the rate is given by

$$\frac{1}{2}E_\theta\left[\log\left\{\sum_{i=1}^{n} j^2\cos^2(j\theta)\right\}\right],$$

which as a function of n is asymptotically of order $\frac{3}{2}\log n$.

The next section considers the rate of learning when the error density belongs to the exponential power family and applies the expected utility criterion to derive optimal sample size calculations for the Michaelis–Menten model with t errors.

3. EXAMPLES

3.1. *Exponential Power Family Errors*

Suppose that the error distribution $f(\ )$ is from the exponential power family (Box and Tiao, 1973). The log-likelihood is given by

$$\log f_i(y_i|\theta) = \log a(\sigma,\alpha) - \frac{1}{2}\left|\frac{y_i-\eta_i}{\sigma}\right|^{\alpha}$$

for a suitable normalizing constant $a(\sigma,\alpha)$. For illustration, suppose that we have replications at a single design point. Then, Le Cam (1986) shows that the appropriate normalization sequence is given by $\psi_n(\theta) = n^{-1/2}$, $\alpha > \frac{1}{2}$, $\psi_n(\theta) = (n\log n)^{-1/2}$, $\alpha = \frac{1}{2}$ and $\psi_n(\theta) = n^{-1/(2\alpha+1)}$, $0<\alpha<\frac{1}{2}$. Hence, the rate of learning, $E_\theta\{\log\psi_n(\theta)\}$, depends on the value of α. Heuristically, for small values of α the density is peaked at the origin resulting in the fact that fewer observations are needed to give the same gain in information as when $\alpha > \frac{1}{2}$.

3.2. *Michaelis–Menten Model with t-errors*

Suppose that observations are generated from a model of the form

$$y_i = \frac{\beta}{\alpha + x_i} + \epsilon_i$$

where ϵ_i are independent and identically distributed with a t-distribution on m degrees of freedom with location 0 and known scale parameter σ. Let $\theta = (\alpha,\beta)$ be the parameter of interest. Now,

$$I_f = \frac{1}{\sigma^2}\left(\frac{m+1}{m+3}\right)$$

and the determinant of Fisher information $|I(\alpha,\beta)| = \alpha^2|h(X,\beta)|$ where

$$h(X,\beta) = \sum_{i=1}^{n}\frac{x_i^2}{(\beta+x_i)^4}\sum_{i=1}^{n}\frac{x_i^2}{(\beta+x_i)^2} - \left\{\sum_{i=1}^{n}\frac{x_i^2}{(\beta+x_i)^3}\right\}^2.$$

Hence the expected utility defined in equation (5) is

$$U^{\Theta}(\mathscr{E}) = E_{\alpha}(\log \alpha) + \frac{1}{2} E_{\beta}\{\log |h(X, \beta)|\} + H\{p(\alpha, \beta)\} + \log\left(\frac{m+1}{m+3}\right) - \log(2\pi e \sigma).$$

The criterion adopted for the choice of n is $U^{\Theta}(\mathscr{E}) = c$, similar to that proposed by Lindley (1956). Here $c = c(n)$ is to be interpreted as representing the relative cost of collecting n observations to learning a bit of information about θ. Logarithms are therefore taken to the base 2 to put the information on the bit scale (Rényi, 1961). Hence, the optimal sample size n^* satisfies

$$E_{\alpha,\beta}\left\{\log \left| \sum_{i=1}^{n^*} I_i(\theta) \right|^{1/2}\right\} + H\{p(\alpha, \beta)\} - \log(2\pi e) = c(n^*).$$

For illustration, suppose that the experimenter samples uniformly over the design points $x_i \in \{-4, -3, -2, -1\}$ and that $m = 4$, $\sigma = 0.1$. Suppose that $p(\alpha, \beta)$ is uniform on $[0, e] \times [-0.5, 0.5]$, so that $E_{\alpha}(\log \alpha) = 0$ and $H\{p(\alpha, \beta)\} = 1$. Let r denote the number of replications at each point, so $n = 4r$. Then under this design and definition (5), $U^{\Theta}(\mathscr{E}) = \frac{1}{2} \log r + 0.38$. Suppose that $c = 2$; then the experimenter selects r such that $U^{\Theta}(\mathscr{E})/\log 2 = c$, i.e. $r^* = 8$.

Finally, a link between the regularity condition (11) of Appendix A and the Jeffreys prior is noted. Let

$$J_{X,n}(\theta) = \left| I_f \sum_{i=1}^{n} \eta_i'(x_i, \theta)^{\mathrm{T}} \eta_i'(x_i, \theta) \right|^{1/2}$$

be the unnormalized Jeffreys prior; then regularity condition (11) becomes,

$$\sup_{\theta, \theta_1 \in \Theta} \left\{\frac{J_{X,n}(\theta)}{J_{X,n}(\theta_1)}\right\} < \infty. \qquad (6)$$

Now, the weak convergence of the likelihood ratio process (3) can be used to prove asymptotic posterior normality (Heyde and Johnstone, 1979; Basawa and Prakasa Rao, 1980). Hence inequality (6) provides a condition, in terms of the Jeffreys prior, on the design and regions of the parameter space to check for asymptotic posterior normality. Hence, from a practical perspective the Jeffreys prior is a useful tool for identifying regions where asymptotic posterior normality might fail. For example, it fails in regions of unidentifiability (Hills, 1987). Heuristically, this is not surprising as it is these regions where the experimenter will not learn from experimentation.

ACKNOWLEDGEMENTS

I would like to thank David Stoffer, Larry Wasserman and the referee for many helpful comments.

APPENDIX A

The Shannon information gain (2) can be expressed in terms of the likelihood ratio process (3) as follows:

$$I^{\Theta}(\mathscr{S}) = \int_{\Theta} \int p(\theta) f(\mathbf{y}\,|\,\theta) \log\left\{\frac{f(\mathbf{y}\,|\,\theta)}{p(\mathbf{y})}\right\} \mathrm{d}\mathbf{y}\, \mathrm{d}\theta,$$

where $p(\mathbf{y}) = \int_{\Theta} f(\mathbf{y}\,|\,\theta) p(\theta)\, \mathrm{d}\theta$. Under the change of variables $\theta \mapsto \theta + \psi_{X,n}(\theta)\alpha$ and by definition of $Z_{n,\theta}(\alpha)$, we have

$$\frac{p(\mathbf{y})}{f(\mathbf{y}\,|\,\theta)} = |\psi_{X,n}(\theta)| \int Z_{n,\theta}(\alpha) p(\theta + \psi_{X,n}(\theta)\alpha)\, \mathrm{d}\alpha.$$

Therefore, the information gain decomposes as

$$I^{\Theta}(\mathscr{S}) = -E_{\theta}\{\log|\psi_{X,n}(\theta)|\} - E_{\theta}\{\log p(\theta)\} - E_{\mathbf{y},\theta}\left[\log\left\{\int \frac{p(\theta + \psi_{X,n}(\theta)\alpha)}{f(\theta)} Z_{n,\theta}(\alpha)\, \mathrm{d}\alpha\right\}\right]. \tag{7}$$

Suppose that $Z_{n,\theta}(\alpha)$ weakly converges to $Z_{\theta}(\alpha)$. Typically, $Z_{\theta}(\alpha) =' \exp(\alpha^{\mathrm{T}} Z - \frac{1}{2}\alpha^{\mathrm{T}}\alpha)$ where Z is a standard multivariate normal random variable (Basawa and Prakasa Rao, 1980). Moreover, suppose that $\psi_{X,n}(\theta) \rightarrow 0$ and that the following interchange of $\lim_{n\rightarrow\infty}$ and integral signs is valid:

$$\lim_{n\rightarrow\infty}\left(E_{\mathbf{y},\theta}\left[\log\left\{\int \frac{p(\theta + \psi_{X,n}(\theta)\alpha)}{p(\theta)} Z_{n,\theta}(\alpha)\, \mathrm{d}\alpha\right\}\right]\right) = E_{Z,\theta}\left[\log\left\{\int Z_{\theta}(\alpha)\, \mathrm{d}\alpha\right\}\right]. \tag{8}$$

Then, by equation (7),

$$\lim_{n\rightarrow\infty} [I^{\Theta}(\mathscr{S}) + E_{\theta}\{\log|\psi_{X,n}(\theta)|\}] = -E_{\theta}\{\log p(\theta)\} - E_{Z,\theta}\left[\log\left\{\int Z_{\theta}(\alpha)\, \mathrm{d}\alpha\right\}\right]. \tag{9}$$

For the non-linear model equation (9) becomes

$$\lim_{n\rightarrow\infty}\left\{I^{\Theta}(\mathscr{S}) - E_{\theta}\left(\log\left|\sum_{i=1}^{n} \frac{\partial\eta_i^{\mathrm{T}}}{\partial\theta}\frac{\partial\eta_i}{\partial\theta}\right|^{1/2}\right)\right\} = H\{p(\theta)\} + \frac{k}{2}\log\left(\frac{I_f}{2\pi e}\right), \tag{10}$$

as required. Therefore, regularity conditions are required to justify step (8). For illustration, assume that Θ is an interval of $\mathbf{R}$. Sufficient regularity conditions (Ibragimov and H'asminsky, 1975a, b; Le Cam, 1986) on the likelihood ratio process are as follows.

Let $\phi_X^{-2}(\theta) = I_f \psi_{X,n}^2(\theta)$ and suppose that $\phi_X^{-2}(\theta_1) < c\,\phi_X^{-2}(\theta_2)$ for some positive constant c and all $\theta_1, \theta_2 \in \Theta$, i.e.

$$\sup_{\theta,\theta_1\in\Theta}\left\{\frac{\sum_{i=1}^{n} \eta_i'(x_i,\theta)^{\mathrm{T}}\eta_i'(x_i,\theta)}{\sum_{i=1}^{n} \eta_i'(x_i,\theta_1)^{\mathrm{T}}\eta_i'(x_i,\theta_1)}\right\} < \infty. \tag{11}$$

Moreover, suppose that the random variables, $\eta_i(\theta) = f_i'(Y_i\,|\,\theta)/f_i(Y_i\,|\,\theta)$ satisfy the Lindeberg condition,

$$\lim_{n\rightarrow\infty}\sup_{\theta\in\Theta}\left[\phi_X^2(\theta)\sum_{i=1}^{n} E_{\theta}\{\eta_i^2(\theta)\, I_{\{|\eta_i(\theta)| < \epsilon\phi_X^{-1}(\theta)\}}\}\right] = 0$$

where I_A is the indicator of the set A. Now, let $\theta_0 \in \Theta$ be an arbitrary fixed value of the parameter. Suppose that there exists δ, $0 < \delta < 1$, such that, as $n \rightarrow \infty$,

$$\gamma(n) = \phi_X^2(\theta_0) \sup_{\theta_2 \in \Theta} \sup_{|\theta_2-\theta_1|<\phi_X^\delta(\theta_0)} \left[\sum_{i=1}^n \int \left\{ \frac{\partial}{\partial\theta} f_i^{1/2}(y|\theta_2) - \frac{\partial}{\partial\theta} f_i^{1/2}(y|\theta_1) \right\} \nu_i(\mathrm{d}y) \right] \rightarrow 0$$

and, for some positive β,

$$\lim_{n\rightarrow\infty} \{\phi_X^\beta(\theta_0)\} \inf_{\theta_i \in \Theta} \inf_{|\theta_2-\theta_1|<\phi_X^\delta(\theta_0)} \left[\sum_{i=1}^n \int \{f_i^{1/2}(y|\theta_2) - f_i^{1/2}(y|\theta_1)\} \nu_i(\mathrm{d}y) \right] > 0.$$

Now, assume that the prior $p(\theta)$ satisfies the smoothness condition $|p(\theta+s) - p(\theta)| < c|s|^{\epsilon_0}$ for some c, ϵ_0, $\sup_\Theta\{p(\theta)\} < \infty$, and the likelihood ratio process is such that $\int_{|\alpha|>A} Z_{n,\theta}(\alpha)\,\mathrm{d}\alpha$ weakly converges to 0 as $A \rightarrow \infty$. Finally, suppose that the prior $p(\theta)$ is such that

$$KL(s) = \int p(\theta) \log\left\{\frac{p(\theta+s)}{p(\theta)}\right\} \mathrm{d}\theta$$

tends to 0 as $s \rightarrow 0$ and satisfies $|KL(s)| < c(1+|s|^p)$, for some c. Then, by a modification of theorem 3.1 of Ibragimov and H'asminsky (1973) the formal interchange in equation (8) is valid.

REFERENCES

Basawa, I. V. and Prakasa Rao, P. L. S. (1980) *Statistical Inference for Stochastic Processes*. London: Academic Press.
Bernardo, J. M. (1979) Reference posterior distributions for Bayesian inference (with discussion). *J. R. Statist. Soc.* B, **41**, 113–147.
Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*. Reading: Addison-Wesley.
Brooks, R. J. (1980) On the relative efficiency of two paired-data experiments. *J. R. Statist. Soc.* B, **42**, 186–191.
——— (1987) Optimal allocation for Bayesian inference about an odds ratio. *Biometrika*, **74**, 196–199.
Chaloner, K. and Larntz, K. (1989) Optimal Bayesian design applied to logistic regression experiments. *J. Statist. Planng Inf.*, **21**, 191–208.
DeGroot, M. H. (1962) Uncertainty, information and sequential experiments. *Ann. Math. Statist.*, **33**, 404–420.
Heyde, C. C. and Johnstone, I. M. (1979) On asymptotic posterior normality for stochastic processes. *J. R. Statist. Soc.* B, **41**, 184–189.
Hills, S. (1987) Reference priors and identifiability problems in nonlinear models. *Statistician*, **36**, 235–240.
Ibragimov, I. A. and H'asminsky, R. Z. (1973) On the information contained in a sample about a parameter. In *Proc. 2nd Int. Symp. Information Theory*, pp. 295–309. Debrecen: Akademiai Kiado, Budapest.
——— (1975a) Local asymptotic normality for non-identically distributed observations. *Theory Probab. Applic.*, **20**, 246–260.
——— (1975b) Properties of maximum likelihood and Bayes estimators for non-identically distributed observations. *Theory Probab. Applic.*, **20**, 689–697.
Kullback, S. (1959) *Information Theory and Statistics*. New York: Wiley.
Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*. New York: Springer.
Lindley, D. V. (1956) On the measure of information provided by an experiment. *Ann. Statist.*, **27**, 986–1005.
Polson, N. G. (1988) Some Bayesian perspectives on statistical modelling. *PhD Thesis*. University of Nottingham.
Rényi, A. (1961) On measures of entropy and information. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 547–561. Berkeley: University of California Press.
Smith, A. F. M. and Verdinelli, I. (1980) A note on Bayes designs for inference using a hierarchical linear model. *Biometrika*, **67**, 613–619.
Stone, M. (1959) Application of a measure of information to the design and comparison of regression experiments. *Ann. Statist.*, **21**, 55–70.