Information Theoretic Asymptotics of Bayes Methods

Bertrand S. Clarke [1]

Andrew R. Barron [2]

Abstract

An information theoretic quantity plays a key role in the application of Bayes methods of parametric density estimation to universal data compression, to composite hypothesis testing, and to stock market portfolio selection.

This quantity is the Kullback - Leibler distance between the true density and the mixture of densities with respect to a prior. It is shown to be approximated by a constant plus one half the logarithm of the total Fisher information, which is close to (d/2)log n, where d is the dimension of the parameter space, and n is the sample size. Consequences for the applications are discussed.

Submitted to the *Transactions on Information Theory*

[1] Clarke is with the Department of Statistics, University of Illinois at Urbana - Champaign.

[2] Barron is with the Department of Statistics, Department of Electrical and Computer Engineering and the Coordinated Science Laboratory at the University of Illinois.

This work was supported in part by the Joint Services Electronics Program, contract number N 00014-84-C-0149.

1. Introduction

The Kullback-Leibler number is a measure of distance between two densities which arises naturally in certain contexts. It can be used as a loss function in estimation, see Aitchison [1], Kullback [17]; as a measure of redundancy in coding problems see Davisson [13]; and, as the error exponent for testing hypotheses see Bahadur [2], Blahut [6]. It defines a mode of convergence which is stronger than L_1 and Hellinger distance, but weaker than the chi-square distance between distributions, Csiszar [12]. Intuitively speaking, it puts a higher proportion of weight on the tails of distributions.

Here we develop an asymptotic expression for the Kullback - Leibler number between frequentist and Bayesian distributions, and apply the result to several contexts.

Let P_θ be a family of probability measures, indexed by a d - dimensional parameter vector θ. Let random variables $X_1, X_2, \ldots, X_n$ be conditionally independent given θ with joint product distribution P_θ^n . A frequentist assumes that there is a fixed, but unknown, true parameter value θ_o, for which the data is governed by $P_{\theta_o}^n$. A Bayesian assumes a prior density function $w(\theta)$, so that the (marginal) distribution for the data is $M_n = \int P_\theta^n \; w(\theta) d\theta$, the mixture of the product distributions obtained by integrating out the parameter.

We examine, in detail, the Kullback - Leibler distance between the joint distributions $P_{\theta_o}^n$ and M_n . It has been shown that under very general conditions, which include infinite dimensional parametric families, that

$$\lim_{n \to \infty} \frac{1}{n} D(P_{\theta_o}^n \; || M_n) = 0,$$

see Barron [3], 4]. In contrast, the distance between any two distinct product measures $(1/n) D(P_{\theta_o}^n \; || P_\theta^n) = D(P_{\theta_o} \; || P_\theta)$ remains fixed away from zero.

Thus, for large sample sizes, the Bayesian distribution M_n , which we know, is not far from the frequentist ‘‘true’’ distribution $P_{\theta_o}^n$, which is unknown.

In this paper we show that for smooth parametric families the Kullback - Leibler number is approximated by one half the logarithm of the Fisher information minus the logarithm of the prior density:

$$D(P_{\theta_o}^n \; || M_n) = \frac{1}{2}\log \; | I_n(\theta_o) | - \log w(\theta_o) - \frac{d}{2}\log 2\pi e + {}_o(1), \tag{1}$$

where ${}_o(1) \to 0$ as $n \to \infty$, see Theorem 4. 1. Here, $I_n(\theta_o) = nI(\theta_o)$ is the total Fisher information in the sample, which has determinant $| I_n(\theta_o) | = n^d \; | I(\theta_o) |$. Thus,

$$D(P_{\theta_o}^n \; || M_n) \sim \frac{d}{2}\log n + c,$$

where $c = \log[(2\pi e)^{\frac{d}{2}} \ | I(\theta_o) \ |^{\frac{1}{2}} / w(\theta_o)]$. So, the divergence of the Bayes and frequentist distributions is precisely characterized. Although $D(P^n_{\theta_o} \ | \ | M_n)$ slowly tends to infinity, the divergence per sample, $(1/n)D(P^n_{\theta_o} \ | \ | M_n)$, is of order (log n)/n. The (d/2)log n term has previously been identified for certain cases in universal source coding by Krichevsky and Trofimov [16], and Rissanen [20].

The asymptotic distribution and the asymptotic expected value of the log density ratio is determined. It is seen that

$$2\left[\ \log \frac{m(X_1,...,X_n)}{p(X_1,...,X_n \ | \theta_o)} + D(P^n_{\theta_o} \ | \ | M_n) \right]$$

converges, in distribution, to $\chi^2_d - d$ where χ^2_d has a chi-square distribution with d degrees of freedom.

After defining some notation in section 2, we discuss implications of the main result in section 3. It is seen that $D(P_\theta{}^n \ | \ | M_n)$ is: (A), the cumulative risk of Bayes estimators of the density function;

(B), the redundancy of a source code based on M_n; (C), the exponent of error probability for Bayes tests of a simple versus composite hypothesis;

and (D), a bound on the financial loss in a stock market portfolio selection problem. In section 4, we formally state the conditons under which we have proved the theorem. In section 5, we state the two approximations used in the proof of the theorem, which is given in section 6. In appendices,

we give the proofs of the two approximations, and prove two lemmata which are used in the proofs of the approximations.

The proof of our main theorem hinges on a technique for approximating integrals first introduced by Laplace in 1774, see Stigler [21], for a special case of the integral $\int p_\theta(x_1, . . . , x_n) w(\theta) d\theta$

. Laplace's approximation for such integrals is now a standard technique in statistical analysis. Walker [25], and Tierney and Kadane [22], provide two examples and some general theory is presented by De Bruijn [14].

2. Notation

Let **(X,B)** be a measurable space and **P**=$\{P_\theta \mid \theta \in \Omega\}$ be a collection of probabilities, each of which is defined on it. These probability measures are assumed to have probability density functions $p_\theta(x)$, with respect to a fixed sigma finite meaure $\lambda(dx)$, which are distinct for distinct θ . Assume the parameter space is contained in some R^d, and is an open set or the closure of an open set. For each natural number n, assume that X_n is a random variable defined on **(X,B)**, taking values x_n , and that $<X_k>_{k=1}^n = X^n$ is a sequence of independently and identically distributed random variables with outcomes denoted x^n.

Given that the true distribution is an element of the parametric family **P**, the problem is to estimate the density from the random sample.

We use the Kullback - Leibler number as an assessment of how much one distributions on a measurable space differ from each other. For densities p, q the Kullback - Leibler number is

$$D(p \mid\mid q)=E_p \log \frac{p(x)}{q(x)},$$

or equivalently denoted by

$$D(P \mid\mid Q).$$

Except where noted otherwise, we use the natural logarithm and denote it log.

If a prior distribution is assumed, then the marginal density function for X^n, with respect to λ^n, is the mixture of the conditional densities $p_n(x^n \mid \theta)=\Pi_1^n p(x_i \mid \theta)$ obtained by integrating with respect to the prior, i.e. ,

$$m_n(X^n)=\int_\Omega w(\theta)p(X^n \mid \theta)d\theta,$$

where w is the density function for the prior with respect to Lebesgue measure on R^d. We denote the mixture distribution itself by M_n, and use the notations $p_\theta(x)$ and $p(x \mid \theta)$ interchangably as convenience dictates. Note that although X^n is a sample of n independently and identically distributed random variables under p_θ^n , under m_n they are no longer independent, in general. However, the dependence on the unknown parameter has been removed. Assume that θ_o is the true value of the parameter. A natural question is: how different is $m_n(X^n)$ from $p(X^n \mid \theta_o)$ i. e., how much accuracy is sacrificed if we model the density by m_n. We answer this question by examining $D(P_{\theta_o}^n \mid\mid M_n)$. Let $\hat{\theta}$ be the maximum likelihood estimate for θ_o, the M.L.E. , and,

$$I(\theta)=E_\theta[\frac{\partial^2}{\partial\theta_j\partial\theta_k}p(X \mid \theta)]_{j,k=1...d}$$

be the Fisher information matrix. We adopt the convention that E denotes expectation with respect to P_{θ_o}, and we omit superscript n's on product measures where the meaning is clear from the context.

3. Applications

A. *Implications for Density Estimation.*

One natural estimator of $p(x \mid \theta)$ at any given x is the mean of the posterior distribution

$$\hat{p}_n(x;X^n)=\int_\Omega p_\theta(x)w(\theta \mid X^n)d\theta.$$

Observe that this estimator is the predictive density

$$\hat{p}_n(x)=m(X_{n+1}=x \mid X^n).$$

Adapting a result due to Aitchison [1], we have the following.

Proposition 3.A: $\hat{p}_n$ is the Bayes estimator of the density function. The cumulative risk of this estimator is

$$\sum_{k=1}^{n} E\ D(p_{\theta_o} \mid\mid \hat{p}_k)=D(P_{\theta_o}^n \mid\mid M_n).$$

Consequently, under the conditions of Theorem 4.1, the cumulative risk is approximated by (d/2) log n + c, and the average risk $(1/n)\sum E\ D(p \mid\mid \hat{p}_k)$ converges to zero at rate (log n)/n .

Proof: The information inequality, $D(p \mid\mid q)\geq 0$, with equality if and only if $p=q$, implies that $\hat{p}_n$ is the Bayes estimator,

since, for any other density q, the posterior average of the risk is seen to equal

$$\int_\Omega D(p_\theta \mid\mid q)w(\theta \mid X^n)d\theta=\int_\Omega w(\theta \mid X^n)D(p_\theta \mid\mid \hat{p}_n)\,d\theta+D(\hat{p}_n \mid\mid q).$$

So, we see that the minimum is achieved when the second term is zero, i. e. , when $q=\hat{p}_n$.

By Bayes rule, $\hat{p}_n$ equals the predictive density, which is

$$m(X_{n+1}=x_{n+1} \mid X^n) = \frac{m_{n+1}(x^{n+1})}{m_n(x^n)}.$$

So, by the chain rule for the Kullback-Liebler, number we have that

$$D(P^n_{\theta_o} \mid\mid M_n) = \sum_{k=1}^{n} E\ D(P_{\theta_o} \mid\mid \hat{P}_k),$$

where each summand is the risk in estimating the density using the Bayes estimate based on k observations. □

We remark that the individual risk terms $E\ D(P_{\theta_o} \mid\mid \hat{P}_n)$ also converge to zero as $n \to 0$. This follows from noting that

$$E\ D(P_{\theta_o} \mid\mid \hat{P}_n) = D(P^n_{\theta_o} \mid\mid M_n) - D(P^{n-1}_{\theta_o} \mid\mid M_{n-1}),$$

and applying the theorem to each term on the right hand side. Thus, the posterior mean density estimator is consistent for the density in expected Kullback - Leibler distance. Note that cumulative risk of the order (d/ 2)log n suggests that the risk $E\ D(P_{\theta_o} \mid\mid \hat{P}_n)$ is of order d/(2n). This same rate, d/(2n), was identified by Ĉencov, [7] pp. 434, for the maximum likelihood density $P_{\hat{\theta}}$.

B. Applications to Universal Source Coding.

Consider the problem of providing a noiseless source code with small expected length for a block of discrete data $X^n=(X_1, \ldots, X_n)$, when the discrete probability density is assumed to be a member of P but otherwise unknown. Many have studied this problem extensively, for instance Davisson [13]. Recall that if

$$\phi: \mathbf{X}^n \to \{0,1\}^*$$

is a uniquely decodeable code with codelengths $l(\phi(X^n))$, where the asterisk indicates the set of all finite length strings of elements of the set, then

$$Q_n(X^n) = 2^{-l(\phi(X^n))}$$

defines a subprobability mass function on $\mathbf{X}^n$, by the Kraft-McMillan inequality. The redundancy of a code $\Phi = \{ \phi(X^n) \mid X^n \in \mathbf{X}^n \}$ is the difference between the expected length of a message and its lower bound, the entropy:

$$R_n(\Phi, P_{\theta_o}) = E\,[l(\phi(X^n)) - \log(\frac{1}{P_{\theta_o}(X^n)})]$$

$$= E\,[\log(\frac{1}{Q^n(X^n)}) - \log(\frac{1}{P_{\theta_o}(X^n)})]$$

$$= D(P^n_{\theta_o} \;||\; Q^n),$$

where log is taken base 2. Thus the redundancy is the Kullback - Leibler number. We want to choose l so as to minimize the redundancy. Among all subprobability mass functions Q, the one which minimizes the average of $D(p^n_\theta \;||\; q^n)$ with respect to a prior w is the mixture m_n . It is well known that, for all p_θ, the Shannon code based on m_n , i.e., the one with code lengths

$$l(\phi(X^n)) = \lceil \log \frac{1}{m_n(X^n)} \rceil ,$$

has redundancy within 1 bit of $D(p^n_{\theta_o} \;||\; m_n)$.

The concepts of noiseless source coding of discrete data may also be applied to the case of continuous random variables which are arbitrarily finely quantized. In the sense made clear by the following proposition, the relative entropy remains the redundancy for nondiscrete sources. If a noiseless code is specified for every finite quantization of a nondiscrete source,

we define the redundancy to be the supremum of the redundancies over all such quantizations.

Proposition 3.B: For a nondiscrete source, the redundancy of the Shannon code based on M_n is $D(P^n_{\theta_o} \;||\; M_n)$, to within one bit.

Thus the redundancy of the Bayes code is given asymptotically by

$$\frac{d}{2}\log\frac{n}{2e\pi} + \frac{1}{2}\log\det I(\theta_o) - \log w(\theta_o),$$

under the conditions of Theorem 4.1.

Proof: For any finite partition π, of $\mathbf{X}^n$, we can specify a code book Φ, by use of the Shannon code based on the probability measure restricted to π . For the Shannon code we have an explicit codelength formula:

$$l(\Phi_{n,\pi}(A)) = \lceil \log \frac{1}{Q_n(A)} \rceil$$

and the redundancy is:

$$R_{\pi,n}(\Phi_n, P_\theta) = \sum_{A\in\pi} l(\phi_n(A))P^n_\theta(A) - P^n_\theta(A)\log(\frac{1}{P^n_\theta(A)}).$$

So, to within one bit, the redundancy on the partition is the discrete divergence $\sum_{A\in\pi} P^n_\theta(A)\log P^n_\theta(A)/Q_n(A)$. Taking the supremum over all possible partitions gives $D(P^n_\theta \mid\mid Q_n)$, by using a well known theorem, see Kullback,

Keegel and Kullback [18], pp. 6-7. If Q_n is replaced by M_n , then we get the Bayes code, and the result is the asymptotic least upper bound on the redundancy. □

Rissanen [20], gave (d/2n)log n as a lower bound on the redundancy, with error O(log n) . Our extension identifies the constant so that we have a better approximation: o(1). The most stringent hypothesis in [20] is that $\hat{\theta}$ be asymptotically normal. Our hypotheses are about as strong. Sufficient conditions for the asymptotic normality of $\hat{\theta}$ are given by Lehmann [19] pp. 429-430, and Cramer [11] pp. 500-501.

While we have assumed a bound on the expected supremum of the squares of the second derivatives, both Lehmann and Cramer assume a bound on the expected supremum of the absolute values of the second and third derivatives.

We have used a higher moment rather than a higher derivative.

C. An Application to Hypothesis Testing.

Consider the hypothesis test H: P_{θ_o} versus K: P_θ, $\theta \neq \theta_o$. We constrain that the probability of type 1 error is not more that $\alpha_1 \in (0, 1)$, and examine the performance of tests in terms of the probability of type 2 error averaged with respect to a prior density $w(\theta)$ over the class of alternatives K. Let $c(\alpha)$ be the $1-\alpha$ quantile of a centered chi-square random variable with d degrees of freedom, i.e., $P(\chi^2_d - E\chi^2_d > c) = \alpha$. The (Bayes) optimal test compares the test statistic $\log m_n(x^n)/p(x^n \mid \theta_o)$ to a critical value $t = t_n(\alpha_1)$. The following proposition shows how to select the critical value in practice. Specifically, Theorem 4.1 gives a convenient approximation to it. Moreover, the average power of the test is shown to be related to D.

Proposition 3.C: The asymptotic level α_1 critical value for the Bayes test is $D(P^n_{\theta_o} \mid\mid M_n) - \frac{1}{2}c(\alpha_1)$ and the optimal average probability of type 2 error is, to within a constant factor dependent only on α_1,

$$\alpha_2 \doteq e^{-D(P^n_{\theta_o} \mid\mid M_n)}$$

$$\doteq \frac{n^{-\frac{d}{2}}(2\pi e)^{\frac{d}{2}} w(\theta_o)}{\sqrt{\det I(\theta_o)}},$$

in the sense that there exists a bounded interval $[\, L(\alpha_1), U(\alpha_1)\,]$ such that every test with type 1 error less than or equal to α_1 satisfies

$$\liminf_{n \to \infty} [\, \log \alpha_2 + D(P^n_{\theta_o} \; | \, | \, M_n)\,] \geq L(\alpha_1),$$

and there exists a test with type 1 error α_1 for which the upper bound

$$\limsup_{n \to \infty} [\, \log \alpha_2 + D(P^n_{\theta_o} \; | \, | \, M_n)\,] \leq U(\alpha_1)$$

holds.

The functions L and U can be expressed in terms of $c(\alpha)$.

Remark 1: This extends Stein's lemma, see Chernoff [9], or Bahadur [2], for simple versus simple hypotheses, say P_{θ_o} versus P_θ
for some $\theta \neq \theta_o$, which asserts that

$$\alpha_2 \doteq e^{-D(P^n_{\theta_o} \; | \, | \, P^n_\theta)} .$$

Remark 2: The classical likelihood ratio test, L.R.T., uses the statistic $\log [\, p(x^n \; | \, \hat{\theta}) / p(x^n \; | \, \theta_o)]$. Proposition 1 relates the likelihood ratio test to the Bayes test: since

$$\log \frac{m_n(X^n)}{p(X^n \; | \, \theta_o)} = \log \frac{p(X^n \; | \, \hat{\theta})}{p(X^n \; | \, \theta_o)} + \log \frac{m_n(X^n)}{p(X^n \; | \, \hat{\theta})}$$

$$\sim \log \frac{p(X^n \; | \, \hat{\theta})}{p(X^n \; | \, \theta_o)} + \frac{d}{2} \log \frac{2\pi}{n} + \log \det I^*(\hat{\theta})^{-1},$$

we see that the L.R.T. and the Bayes test are asymptotically equivalent,
a fact which has been previously observed in specific cases. Moreover,

$$2 \log \frac{p(X^n \; | \, \hat{\theta})}{p(X^n \; | \, \theta_o)}$$

has an asymptotic chi-square distribution with d degrees of freedom, see Wilks [26].

Proof: First we prove the lower bound statement. Let $\tilde{C}_n$ be any critical region with $P_{\theta_o}(\tilde{C}_n) \leq \alpha_1$, and let A_n be the "typical set"

$$A_n = \{\, x^n \; | \log \frac{p(x^n \; | \, \theta_o)}{m_n(x^n)} \leq D(P^n_{\theta_o} \; | \, | \, M_n) - \frac{1}{2} c(\alpha) \}$$

where $\alpha > \alpha_1$. Observe that

$$\lim_{n \to \infty} P^n_{\theta_o}(A_n) = \alpha.$$

Then the average probability of type 2 error satisfies

$$\alpha_2 = M_n(\tilde{C}^c_n) \geq M_n(\tilde{C}^c_n \cap A_n) \geq e^{-D(P^n_{\theta_o} \; || M_n) + \frac{1}{2}c(\alpha)} P^n_{\theta_o}(\tilde{C}^c_n \; (ca \; A_n)$$

$$\geq e^{-D(P^n_{\theta_o} \; || M_n) + \frac{1}{2}c((*a)} [\, P^n_{\theta_o}(\tilde{C}^c_n) - P^n_{\theta_o}(A^c_n)].$$

Since

$$\lim_{n \to \infty} [\, P^n_{\theta_o}(\tilde{C}^c_n) - P^n_{\theta_o}(A^c_n) \,] = \alpha - \alpha_1 > 0,$$

we may take logarithms to obtain

$$\liminf_{n \to \infty} [\, \log \alpha_2 + D(P^n_{\theta_o} \; || M_n) \,] \geq \frac{1}{2}c(\alpha) + \log(\alpha - \alpha_1).$$

where $\alpha \in (\alpha_1, 1)$. Note that c is strictly decreasing in α and ranges from $-E\chi^2_d$ to ∞ and $\log(\alpha - \alpha_1)$ is strictly increasing. It is possible to get an implicit algebraic relation which must be satisfied by the α which maximizes the right hand side. In particular, we chose $\alpha = (\alpha_1 + 1)/2$ so as to get a lower bound of the form claimed.

Now we prove the upper bound. The Bayes optimal test is of the form reject H if and only if $(X_1, \ldots, X_n) \in C_n$, where C_n is the critical set

$$C_n = \{ \, x^n \; | \log \frac{p(x^n \; | \theta_o)}{m_n(x^n)} \leq t \}.$$

Choosing

$$t = D(P^n_{\theta_o} \; || M_n) - \frac{c(\alpha_1)}{2}$$

we have that

$$-2[\log \frac{p(x^n \; | \theta_o)}{m_n(x^n)} - D(P^n_{\theta_o} \; || M_n)]$$

converges weakly to a chi-square random variable with d degrees of freedom.

So, the limiting probability of type 1 error is

$$\lim_{n \to \infty} P_{\theta_o}(C_n) = \alpha_1.$$

By Markov's inequality, the average probability of type 2 error satisfies

$$\alpha_2 = M_n(C^c_n) \leq e^{-t} = e^{-D(P^n_{\theta_o} \; || M_n) + \frac{1}{2}c(\alpha_1)}.$$

Thus, taking logs, and rearranging gives

$$\limsup_{n \to \infty} [\log \alpha_2 + D(P^n_{\theta_o} \ | \ | M_n)] \le \frac{1}{2} c(\alpha_1),$$

so that $c(\alpha_1)/2$ upper bounds the limit superior of the left hand side. □

D. Application to Portfolio Selection Theory.

Let $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n, \ldots$ be a sequence of stock market return vectors, where the coordinates X_{ij} denote the multiplicative factor by which dollars invested in stock j, j=1,..., k, are increased/decreased during the i^{th} day, or other investment period. At the beginning of each investment period stocks are bought or sold so as to result in a portfolio of stock proportions $\underline{b} = (b_1, \ldots, b_k)$, $b_j \ge 0$, $\sum_{j=1}^{k} b_j = 1$. If the true distribution P_{θ_o} were known, then the portfolio $\underline{b} = \underline{b}(P_{\theta_o})$ would be chosen to acheive

$$W^* = \max_{\underline{b}} E \ \log \underline{b}^T \underline{X}$$

so as to achieve maximum possible exponential growth rate of wealth, see Kelly [15]. Not knowing the true distribution, we may base our portfolio $b_n = b_n(\hat{P}_{n-1})$ for day n on an estimate $\hat{P}_n$ of the true distribution. Barron and Cover [5], have shown that the resulting drop in exponential growth of wealth is bounded by

$$\frac{1}{n} \sum_{i=1}^{n} E \ D(P_{\theta_o} \ | \ | \hat{P}_{i-1}).$$

In particular, if we use the predictive density estimate, $\hat{p}_n(x) = m(x_{n+1} = x \ | x^n)$ then the bound on the decrement is precisely $(1/n) D(P^n_{\theta_o} \ | \ | M_n)$, the very quantity approximated by our theorem. The Bayes sequential investment strategy, which uses the predictive density to select the portfolio, is optimal with respect to M_n. If P_{θ_o} were known, the resulting optimal wealth is

$$S_n^* = e^{n(W^* + {}_o(1))},$$

where ${}_o(1) \to 0$ in probability. We can lower bound the wealth of the Bayes strategy in terms of the optimal wealth.

Proposition 3.D: The Bayes strategy, investing based on M_n , achieves wealth at least

$$S_n \ \dot{\ge} \ S_n^* e^{-D(P_{\theta_o} \ | \ | M_n)}$$

$$\doteq S_n^* (2\pi e n)^{-\frac{d}{2}} \frac{w(\theta_o)}{\sqrt{\det I(\theta_o)}},$$

where the last expression holds asymptotically under the conditions of Theorem 4.1. Indeed, for any $\alpha \in (0, 1)$ and any $\tau > 0$,

$$S_n \geq S_n^* e^{-D(P_{\theta_o}^n \mid\mid M_n) - \frac{1}{2}c(\alpha) - \tau}$$

except on a set with probability asymptotically less than or equal to $(*a + e^{-\tau}$, as $n \to \infty$, where $c(\alpha)$ is the same as in the last proposition.

Proof: By Markov's inequality, the wealth satisfies

$$S_n \geq S_n^* \frac{m_n(X^n)}{P(X^n \mid \theta_o)},$$

except on a set of probability

$$P_{\theta_o}^n (\{ \frac{S_n^*}{S_n} \frac{m_n(X^n)}{p(X^n \mid \theta_o)} \geq e^{\tau} \}) \leq e^{-\tau} E_{\theta_o} \frac{S_n^*}{S_n} \frac{m_n(X^n)}{p(X^n \mid \theta_o)}$$

$$\leq e^{-\tau} E_{m_n} \frac{S_n^*}{S_n}$$

$$\leq e^{-\tau},$$

where the inequality $E_m S_n^*/S_n \leq 1$ follows from the Kuhn - Tucker conditions for the optimality of S_n for the distribution M_n , see [5]. The result then follows as in the proof of the proposition on hypothesis testing from the fact that twice log $m_n(X^n)/p_{\theta_o}(X^n) + D(P_{\theta_o}^n \mid\mid M_n)$, asymptotically,

has a centered chi-square distribution with d degrees of freedom. □

4. Statement of Conditions and of the Main Result

An estimator $\hat{\theta}$ is consistent for θ_o if and only if for any $\varepsilon > 0$, $P_{\theta_o}(\mid\mid \hat{\theta} - \theta_o \mid\mid > \varepsilon)$ tends to zero as n tends to infinity. We will be requiring consistency of the M.L.E. , and of estimators of the Fisher information matrix at rate O(1/n). So, we first identify several assumptions which imply that, and recur through many of the results. The first three parallel Wald's conditions for consistency, see [24], but are stronger, in that they are second moments so as to get the desired rate.

Assumption 1: For each x, as $||\theta||$ increases, $p(x \mid \theta)(\to 0$.

This assumption is convenient so that, if the parameter space does not have compact closure, then we still have that, for any fixed θ_o,

$$\inf_{\{\theta:\ ||\theta_o-\theta||>\varepsilon\}} D(P_{\theta_o} \,||\, P_\theta)$$

is bounded away from zero, which ensures that the only $P_\theta's$ that are close to P_{θ_o} are those for which θ is close to θ_o . Assumption 1 implies the infimum above is positive: By an Egoroff's theorem argument, one can get a strictly positive lower bound for the L^1 distance, and so a strictly positive lower bound for the Kullback-Leibler number.

Assumption 2: For some large r we have that

$$E[\log \sup_{\{\theta':\ ||\theta'-\theta_o||>r\}} \frac{p(X \mid \theta')}{p(X \mid \theta_o)}]^2 <\infty.$$

Assumption 3: For each θ, and for any $\delta > 0$ small enough, the function

$$\rho(x \mid \theta, \delta) = \sup_{\{\theta':\ ||\theta-\theta'||<\delta\}} p(x \mid \theta')$$

satisfies

$$E\ [\log \frac{p(X \mid \theta_o)}{\rho(X \mid \theta, \delta)}]^2 < \infty.$$

Assumption 4: For each x, $p(x \mid \theta)$ is twice continuously differentiable with respect to θ .

This is a necessary assumption for the Fisher information to exist.

Assumption 5: The prior density w on R^d is continuous, and $w(\theta_o)>0$.

Assumption 6: The matrix $I(\theta_o)$ exists, and is positive definite.

We use this assumption so that all the eigenvalues will be positive and the inverse will exist.

Assumption 7: For some $\xi > 0$, we have that

$$E \sup_{\{|\theta-\theta_o|<\xi\}} | \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X_1 \mid \theta) |^2 <\infty.$$

In section 6 we give a proof of the main result of this paper which is an asymptotic expansion for $D(P^n_{\theta_o} \,||\, M_n)$. Our result is the following theorem.

Theorem 4.1: Suppose assumptions 1 through 7 are satisfied. Then

$$D(P^n_{\theta_o} \;||\; M_n) = \frac{d}{2}\log\frac{n}{2e\pi} + \frac{1}{2}\log\det I(\theta_o) - \log w(\theta_o) + o(1),$$

where $o(1) \to 0$ as $n \to \infty$. Moreover,

$$\log\frac{m_n(X^n)}{p(X^n \;|\; \theta_o)} + D(P^n_{\theta_o} \;||\; M_n) \to \frac{1}{2}(\chi^2_d - d)$$

in distribution, where χ^2_d is a chi-square distribution with d degrees of freedom.

Proof: Deferred until section 6.

Remark: The proof we give in section 6 is one of two that we have. The other, while shorter and more sophisticated, seems less intuitive.

The same result holds in probability with the quantities on the right, $I(\theta_o)$ and $w(\theta_o)$, replaced by their estimates $I^*(\hat{\theta})$ and $w(\hat{\theta})$.

5. Two Approximations

Note that the integrand of $D(P^n_{\theta_o} \;||\; M_n)$ is

$$\log\frac{p(X^n \;|\theta_o)}{m_n(X^n)} = \log\frac{p(X^n \;|\theta_o)}{p(X^n \;|\hat{\theta})} + \log\frac{p(X^n \;|\hat{\theta})}{m_n(X^n)}.$$

The first proposition is an approximation to the first term; the second proposition is an approximation to the expected value of the second. We prove both of the propositions using the estimates I^* and $\hat{\theta}$ since we can know them; if they are replaced by their true values the results remain true. They also remain true with the maximum posterior likelihood estimator, the M.P.L.E., used in place of the M.L.E.

Proposition 5.1: Let assumptions 1 through 6 be satisfied. Then, as $n \to \infty$,

$$|\log\frac{p(x^n \;|\hat{\theta})w(\hat{\theta})}{m_n(x^n)} - \frac{d}{2}\log\frac{n}{2\pi} - \frac{1}{2}\log\det I^*(\hat{\theta}) \;| \to 0,$$

in P_{θ_o} - probability, where

$$I^*(\theta) = \left[\frac{1}{n}\sum_{l=1}^{n}\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p(X_l \;|\; \theta)\right]_{i,j = 1,\cdots,n}.$$

Proof: See appendix A.

Remark 1: This result is substantially due to Walker [25].

In the lower bound on $D(P^n_{\theta_o} \mid\mid M_n)$, it will be necessary to have a rate for the convergence in proposition 5.1. We give it as a corollary.

Corollary 5.1: For any $\varepsilon > 0$, if assumption 7 is satisfied also, then the probability of

$$G_n^c = \{ x^n \mid \mid \log \frac{p(x^n \mid \hat{\theta}) w(\hat{\theta})}{m_n(x^n)} - \frac{d}{2} \log \frac{n}{2\pi} - \frac{1}{2} \log \det I^*(\hat{\theta}) \mid > \varepsilon \}.$$

is of order O(1/n).

Proof: See appendix A.

Next, for $\varepsilon > 0$ we will approximate the expected value of

$$\log \frac{p(X^n \mid \theta_o)}{p(X^n \mid \hat{\theta})} \chi_{\Omega_{\varepsilon,n}},$$

where $\Omega_{\varepsilon,n}$ is the set defined by

$$\Omega_{\varepsilon,n} \equiv \{ x^n \in \mathbf{R}^n \mid \mid\mid \hat{\theta}(x^n) - \theta_o \mid\mid < \varepsilon, \sup_{\theta \in B_{\varepsilon'}} \mid\mid I^*(\theta) - I(\theta_o) \mid\mid < \varepsilon \}.$$

in which ε' is less than some value of ξ which satisfies assumption 7, and ε is chosen so small that the every element of the set

$$B(I(\theta_o), \varepsilon) = \{ M \in \mathbf{M}_{d\times d} \mid \mid\mid M - I(\theta_o) \mid\mid < \varepsilon \} \subset \mathbf{R}^{m^2}$$

is invertible, where $\mathbf{M}_{d\times d}$ is the collection of all $d\times d$ matrices. Note that by lemma C.1, $P(\Omega^c_{\varepsilon,n} \mid \theta_o) = O(1/n)$. Now, we state the second proposition:

Proposition 5.2: Assume 1, 2, 3, 4 and 6 are satisfied. Then, if $\nabla \log p(X \mid \theta_o)$ has a finite second moment, as $n \rightarrow$ *(if*

$$\int_{\Omega_{\varepsilon,n}} p(x^n \mid \theta_o) \log \frac{p(x^n \mid \theta_o)}{p(x^n \mid \hat{\theta})} \lambda(dx^n) \rightarrow (mi \frac{d}{2},$$

and

$$\log \frac{p(X^n \mid \hat{\theta})}{p(X^n \mid \theta_o)} \rightarrow \frac{1}{2} \chi_d^2$$

in distribution.

Proof: See appendix B.

Remark 1: The choice of domain of integration is motivated by the desire to force the expectation of $I^*(\theta^{**})^{-1}$ to converge to $I(\theta_o)^{-1}$, where θ^{**} is any point in the parameter space on the line joining θ_o to $\hat{\theta}$. For, on the domain of integration not only are the entries of the approximation closer than ε to their true values, they are close enough to force boundedness of the inverse of the matrix they form.

Remark 3: Wilks [26], Wald [23], and Chernoff [8], established that $2 \log p(X^n \mid \hat{\theta})/p(X^n \mid \theta_o)$ converges in distribution to a chi-squared random variable with d degrees of freedom, (under different hypotheses). However, we are not aware of any proof that the limit of the expectations is the expectation of the limiting chi squared random variable.

6. Proof of the Main Theorem

We will sandwich the desired quantity, $D(P^n_{\theta_o} \mid\mid M_n)$, between upper and lower bounds which will both converge to the same expression. By definition

$$D(P^n_{\theta_o} \mid\mid M_n) = \int_{\mathbf{R}^n} p(x^n \mid \theta_o) \log \frac{p(x^n \mid \theta_o)}{m_n(x^n)} \lambda(dx^n).$$

We decompose the integral into a sum of three terms. The domain of integration for two of the integrals is a suitably restricted subset of the sample space;
it was chosen so that local theory will apply. The third integral is over the complement of the subset and we will prove that it is negligible. Our decomposition is:

$$D(P^n_{\theta_o} \mid| M_n) = \int_{\Omega_{\varepsilon,n}} p(x^n \mid \theta_o) \log \frac{p(x^n \mid \theta_o)}{p(x^n \mid \hat{\theta})} \lambda(dx^n) \qquad (2)$$

$$+ \int_{\Omega_{\varepsilon,n}} p(x^n \mid \theta_o) \log \frac{p(x^n \mid \hat{\theta})}{m_n(x^n)} \lambda(dx^n) \qquad (3)$$

$$+ \int_{\Omega^c_{\varepsilon,n}} p(x^n \mid \theta_o) \log \frac{p(x^n \mid \theta_o)}{m_n(x^n)} \lambda(dx^n). \qquad (4)$$

We will upper and lower bound the three integrals above. First note that by proposition 5.2, for any preassigned $\eta > 0$ the value of (2) is in the interval $(-\frac{d}{2} - \eta, -\frac{d}{2} + \eta)$ for n large enough. Next, we obtain good lower bounds on (3) and (4).

For (3) we will use proposition 5.1 and corollary 5.1. Both results continue to hold if the true Fisher information matrix is used in place of its estimate.

So, we redefine the collection of ‘good’ $x^{n}{}'s$ to be:

$$G_n = \{x^n \ \mid \ \mid \log \frac{p(x^n \mid \hat{\theta})}{m_n(x^n)} + \log W(\theta_o) - \frac{d}{2}\log\frac{n}{2\pi} - \frac{1}{2}\log \det I(\theta_o) \ \mid < \varepsilon\}$$

and write (3) as

$$\int_{\Omega_{\varepsilon,n}\cap G_n} p(x^n \mid \theta_o)\log\frac{p(x^n \mid \hat{\theta})}{m_n(x^n)}\lambda(dx^n) + \int_{\Omega_{\varepsilon,n}\cap G_n^c} p(x^n \mid \theta_o)\log\frac{p(x^n \mid \hat{\theta})}{m_n(x^n)}\lambda(dx^n)$$

$$\geq [\ \frac{d}{2}\log\frac{n}{2\pi} + \frac{1}{2}\log \det I(\theta_o) - \log w(\theta_o) - \varepsilon]\ P_{\theta_o}(\Omega_{\varepsilon,n}\cap G_n)$$

$$-P_{\theta_o}(\Omega_{\varepsilon,n}\cap G_n^c)\int_{\Omega_{\varepsilon,n}\cap G_n^c}\frac{p(x^n \mid \theta_o)}{P_{\theta_o}(\Omega_{\varepsilon,n}\cap G_n^c)}\log\frac{m_n(x^n)}{p(x^n \mid \theta_o)}\lambda(dx^n)$$

which by Jensen’s inequality is

$$\geq -\varepsilon + [\frac{d}{2}\log\frac{n}{2\pi} + \frac{1}{2}\log \det I(\theta_o) - \log w(\theta_o)]P_{\theta_o}(\Omega_{\varepsilon,n}\cap G_n)$$

$$- P_{\theta_o}(\Omega_{\varepsilon,n}\cap G_n^c)\log\left[\ \int_{\Omega_{\varepsilon,n}\cap G_n^c}\frac{m_n(x^n)}{P_{\theta_o}(\Omega_{\varepsilon,n}\cap G_n^c)}\lambda(dx^n)\right] \ . \tag{5}$$

Writing $B_n = \Omega_{\varepsilon,n}\cap G_n^c$, the factor in large brackets,

$$-P_{\theta_o}(B_n)\log M(B_n) + P_{\theta_o}(B_n)\log P_{\theta_o}(B_n),$$

is bounded below by o(1), since $-\log M(B_n) > 0$ and $P_{\theta_o}(B_n) \to 0$. Thus we have the new lower bound:

$$-\varepsilon + [\frac{d}{2}\log\frac{n}{2\pi} + \frac{1}{2}\log \det I(\theta_o) - \log w(\theta_o)]P_{\theta_o}(\Omega_{\varepsilon,n}\cap G_n) - \eta, \tag{6}$$

valid for all large n.

For (4), we will use a Jensen’s inequality argument to show that it is greater than or equal to $-\eta$, for all large n. Indeed, (4) equals

$$-P(\Omega_{\varepsilon,n}^c \ \mid \theta_o)\int_{\Omega_{\varepsilon,n}^c}\frac{p(x^n \mid \theta_o)}{P(\Omega_{\varepsilon,n}^c \mid \theta_o)}\log\frac{m_n(x^n)}{p(x^n \mid \theta_o)}\lambda(dx^n)$$

$$\geq -P(\Omega_{\varepsilon,n}^c \ \mid \theta)\log\int_{\Omega_{\varepsilon,n}}\frac{m_n(x^n)}{P(\Omega_{\varepsilon;n}{}^c \mid \theta_o)}\lambda(dx^n)$$

$$= -P(\Omega_{\varepsilon,n}^c \ \mid \theta_o)\log\frac{M_n(\Omega_{\varepsilon,n}^c)}{P(\Omega_{\varepsilon;n}{}^c \mid \theta_o)}$$

$$\geq P(\Omega^c_{\varepsilon,n} \mid \theta_o) \log P(\Omega^c_{\varepsilon,n} \mid \theta_o)$$

$$\geq -\eta \tag{7}$$

for n large enough, since $P(\Omega^c_{\varepsilon,n} \mid \theta_o) \to 0$.

We now have good lower bounds. It remains to upper bound (3) and (4).

To get a good upper bound on (3), note that (3) is upper bounded by

$$-\int_{\Omega_{\varepsilon,n}} \log \int_{B^*_\delta} e^{-\log \frac{p(x^n \mid \hat{\theta})}{p(x^n \mid \theta)}} w(\theta) d\theta p(x^n \mid \theta_o) \lambda(dx^n),$$

in which we have used

$$B^*_\delta = \{ \theta \mid (\theta - \hat{\theta})^t I^*(\theta^{**})(\theta - \hat{\theta}) < \delta, w(\theta) \geq w(\theta_o)(1 - \delta)\},$$

where θ^{**} is as in Appendix B. Because the domains of integration have been cut down appropriately, we can use a Taylor expansion in the inner logarithm. The last integral equals

$$-\int_{\Omega_{\varepsilon,n}} \log \int_{B^*_\delta} e^{-\frac{n}{2}(\theta - \hat{\theta})^t I^*(\theta^{*}{*})(\theta - \hat{\theta})} w(\theta) d\theta p(x^n \mid \theta_o) \lambda(dx^n). \tag{8}$$

Laplace integration gives a lower bound for the inner integral:

$$\int_{B^*_\delta} e^{-\frac{n}{2}(\theta - \hat{\theta})^t I^*(\theta^{**})(\theta - \hat{\theta})} w(\theta) d\theta$$

$$\geq (1 - 2^{d/2} e^{-n\varepsilon/4})(2\pi)^{d/2} \det[nI^*(\theta^{**})]^{-1/2} w(\theta_o)(1 - \delta). \tag{9}$$

Using (9) in (8) gives the upper bound

$$-\int_{\Omega_{\varepsilon,n}} \log (2\pi)^{d/2} \det[nI^*(\theta^{**})]^{-1/2} p(x^n \mid \theta_o) \lambda(dx^n)$$

$$- P(\Omega_{\varepsilon,n} \mid \theta_o) [\log [1 - 2^{d/2} e^{-n\varepsilon/4}] + \log w(\theta_o) + \log (1 - \delta)],$$

which, by lemmata 1 and 2, gives, upon rearrangement, that (3) is upper bounded by

$$\frac{d}{2} \log (\frac{n}{2\pi}) + \frac{1}{2} \log \det I(\theta_o) + \eta(\delta, \varepsilon) - \log w(\theta_o), \tag{10}$$

in which η tends to zero as δ and ε tend to zero, provided that the limit as n increases has already been taken.

The last quantity to upper bound is (4). As with the upper bound on (3) , we may invert the arguement of the log, restrict the domain of integration in the definition of m, and rewite the inner integrand so that we have an upper bound on (5) which is of the form:

$$-\int_{\Omega^c_{\varepsilon,n}} p(x^n \mid \theta_o) \log \int_{\{\theta : \ \| \theta - \theta_o \| \leq \delta\}} e^{-\log \frac{p(x^n \mid \theta_o)}{p(x^n \mid \theta)}} w(\theta) d\theta \lambda(dx^n).$$

Since θ is restricted to a neighbourhood about θ_o , we can use a Taylor expansion of $\log p(x^n \mid \theta)$:

$$\log p(x^n \mid \theta) - \log p(x^n \mid \theta_o) = \sqrt{n}\,(\theta-\theta_o)^t S_n(\tilde{\theta}),$$

where $S_n(\theta) = (1/\sqrt{n}\,)\nabla \log p(X^n \mid \theta)$. Now, the last integral is

$$-\int_{\Omega_{\varepsilon,n}^c} p(x^n \mid \theta_o) \log \int_{\{\theta:\ \| \theta-\theta_o \| \leq \delta\}} e^{-\sqrt{n}\,(\theta-\theta_o)^t S_n(\tilde{\theta})} w(\theta) d\theta \lambda(dx^n) \tag{11}$$

We can now upper bound (11) by bounding the exponent with

$$(\theta-\theta_o)^t S_n(\tilde{\theta}) < \delta \sup_{\{\theta \mid \| \theta-\theta_o \| \leq \delta\}} \| S_n(\theta) \|.$$

Doing so gives

$$-P(\Omega_{\varepsilon,n}^c \mid \theta_o) \log W(\{\theta:\ \| \theta-\theta_o \| \leq \delta\})$$
$$+ \sqrt{n}\,\delta \int_{\Omega_{\varepsilon,n}^c} \sup_{\{\theta \mid \| \theta-\theta_o \| \leq \delta\}} \| S_n(\theta) \| \; p(x^n \mid \theta_o) \lambda(dx^n), \tag{12}$$

as an upper bound for (4). By consistency the first term is no problem.

For the second term use the Cauchy-Schwartz inequality, lemma C.1, and the fact that, by assmption 7,

$$E \sup_{|\theta-\theta_o| \leq \delta} \| \nabla \log p(X_1 \mid \theta) \|^2 < \infty.$$

Putting the bounds (6), (7), (10) and (12) together we have

$$-\frac{d}{2} - \varepsilon + [\frac{d}{2}\log\frac{n}{2\pi} + \frac{1}{2}\log \det I(\theta_o)] P_{\theta_o}(\Omega_{\varepsilon,n} \cap G_n) - 4\eta + \log\frac{1}{w(\theta_o)}$$
$$\leq D(P_{\theta_o}^n \,\|\, M_n) \tag{13}$$
$$\leq -\frac{d}{2} + \frac{d}{2}\log\frac{n}{2\pi} + \frac{1}{2}\log \det I(\theta_o) + 4\eta + \log\frac{1}{w(\theta_o)}$$

The first conclusion of the theorem now follows since $P_{\theta_o}((\Omega_{\varepsilon,n} \cap G_n)^c) = o(1/n)$.

Finally, by writing the decomposition

$$\log \frac{m(X^n)}{p(X^n \mid \theta_o)} + D(P_{\theta_o}^n \,\|\, M_n)$$
$$= \log \frac{m(X^n)}{p(X^n \mid \hat{\theta})} + \log \frac{p(X^n \mid \hat{\theta}}{p(X^n \mid \theta_o)} + D(P_{\theta_o}^n \,\|\, M_n),$$

we have that proposition 5.2 implies that the middle term goes to $(1/2)\chi_d^2$ in law; and that, by proposition 5.1, and the first part of the theorem the sum of the first term and last term goes to -(d/ 2), in probability. This concludes the proof of the theorem. □

Appendix A

In this appendix we give the proofs of proposition 5.1 and corollary 5.1

Proof of Proposition 5.1 and Corollary 5.1: We indicate how to modify the technique used by Walker. Since we are using estimates wherever possible we must be careful about the errors introduced by the approximations so as to be able to prove the rate stated in the corollary. Consider small sets about θ_o, of the form

$$B^*_\delta = \{ \theta \in R^d \mid (\theta - \hat{\theta})^T I^*(\hat{\theta})(\theta - \hat{\theta}) < \delta \},$$

and

$$B_\delta = \{ \theta \in R^d \mid (\theta - \theta_o)^T I(\theta_o)(\theta - \theta_o) < \delta \}.$$

Now:

$$m_n(X^n) = \int_{B^*_\delta} w(\theta) p(X^n \mid \theta) d\theta + \int_{B^{*c}_\delta} w(\theta) p(X^n \mid \theta) d\theta$$

$$\equiv J_1 + J_2. \tag{1}$$

We want to approximate m_n by J_1. So, we must show that the contribution from J_2 is so small that it can be neglected. To do this, we will first show that J_2 is upper bounded by $p(X^n \mid \hat{\theta})$ weighted by an exponentially small factor. Then, we will get bounds for J_1 in terms of $p(X^n \mid \hat{\theta})$ weighted by a polynomially small factor.

By lemma C.2(i), and C.1 (ii), for any positive $\delta_1 < \delta$
there exists an $\varepsilon > 0$ such that

$$0 \le J_2 \le \int_{B^c_{\delta_1}} w(\theta) p(X^n \mid \theta) d\theta \tag{2}$$

$$\le e^{-n\varepsilon} p(X^n \mid \theta_o) \le e^{-n\varepsilon} p(X^n \mid \hat{\theta}),$$

with P_{θ_o} probability at least $1 - c/n$, for some c, and for all large n.

By using lemma C.2, parts (i) and (ii), we choose $\delta_1 < \delta < \delta_2$ and simplify the domain of integration in J_1:

$$\int_{B_{\delta_1}} w(\theta) p(X^n \mid \theta)\, d\theta \le J_1 \le \int_{B_{\delta_2}} w(\theta) p(X^n \mid \theta)\, d\theta,$$

with probability at least $1 - c/n$ for n large enough. We can use a second order Taylor expansion of $\log p(X^n \mid \theta)$ at $\hat{\theta}$. For some $\theta^{**}(X^n) \in [\theta_o, \hat{\theta}]$ we have

$$p(X^n \mid \theta) = p(X^n \mid \hat{\theta}) e^{-\frac{n}{2}(\theta - \hat{\theta})^t I^*(\theta^{**})(\theta - \hat{\theta})}.$$

By lemmata C.1 and C.2, given $\tau > 0$, we have that for all large n the following three properties hold with probability at least $1 - c/n$: $I^*(\hat{\theta})$ is positive definite, $\hat{\theta} \in B_{(*d_1}$ and

$$(1 - \tau)(\theta - \hat{\theta})^t I^*(\hat{\theta})(\theta - \hat{\theta}) \le (\theta - \hat{\theta})^t I^*(\theta^{**})(\theta - \hat{\theta}) \le (1 + \tau)(\theta - \hat{\theta})^t I^*(\hat{\theta})(\theta - \hat{\theta}).$$

Whenever those three conditions are satisfied, we may use Laplace's method to obtain bounds on J_1 of the following form: for some $\varepsilon' > 0$.

$$p(x^n \mid \hat{\theta}) w(\hat{\theta})(1-\varepsilon'')(\frac{2\pi}{n(1+\tau)})^{\frac{d}{2}} \det I^*(\hat{\theta})^{-\frac{1}{2}}(1-e^{-n\varepsilon'})$$

$$\leq J_1 \tag{3}$$

$$\leq p(x^n \mid \hat{\theta}) w(\hat{\theta})(1+\varepsilon'')(\frac{2\pi}{n(1-\tau)})^{\frac{d}{2}} \det I^*(\hat{\theta})^{-\frac{1}{2}}.$$

This holds with probability at least $1 - c/n$. Note that, by the continuity of w, we can let ε'' tend to zero as δ tends to zero. Now we assemble (2) and (3) so that we have bounds on $m_n(x^n)$:

$$p(x^n \mid \hat{\theta}) w(\hat{\theta})(1-\varepsilon'')(\frac{2\pi}{n(1+\tau)})^{\frac{d}{2}} \det I^*(\hat{\theta})^{-\frac{1}{2}}(1-e^{-n\varepsilon'})$$

$$\leq m_n(x^n) \tag{4}$$

$$\leq p(x^n \mid \hat{\theta}) w(\hat{\theta})(1+\varepsilon'')(\frac{2\pi}{n(1-\tau)})^{\frac{d}{2}} \det I^*(\hat{\theta})^{-\frac{1}{2}}(1+e^{-n\varepsilon/2}).$$

We next identify exactly the quantity that we hope is small. Let

$$\Delta_n = \log m(x^n) - \log [w(\hat{\theta})(\frac{2\pi}{n})^{\frac{d}{2}} \det I^*(\hat{\theta})^{-\frac{1}{2}} p(x^n \mid \hat{\theta})].$$

Equation (4), which gives bounds on $m_n(x^n)$, can be used to show that the limit of Δ_n as $n \to \infty$ exists and is zero by taking logs and rearranging the result so as to get upper and lower bounds on Δ_n which go to zero. □

Appendix B

In this appendix we prove the second proposition stated in section 5.

Proof of Proposition 5.2: We show that

$$\int_{\Omega_{\varepsilon,n}} \log \frac{p(x^n \mid \theta_o)}{p(x^n \mid \hat{\theta})} p(x^n \mid \theta_o) \lambda(dx^n) \to -\frac{d}{2}$$

an n → ∞. Since the integral is over a small set we can again use a Taylor expansion and bound the error. First we expand $\log p(x^n \mid \theta)$ to second order about $\hat{\theta}$. Let

$$Z_n = -\sqrt{n}(\theta_o - \hat{\theta}),$$

then the quantity to be asymptotically approximated by the proposition is

$$-(\frac{1}{2}) \int_{\Omega_{\varepsilon,n}} Z_n^T I^*{}_n(\theta^{**}) Z_n p(x^n \mid \theta_o) dx^n \tag{5}$$

where $\theta^{**} = \theta^{**}(x^n)$ lies on the straight line joining θ_o and $\hat{\theta}$. By a first order application of Taylor's theorem we may write

$$\nabla \log p(x^n \mid \hat{\theta}) - \nabla \log p(x^n \mid \theta_o) = nI*_n(\tilde{\theta})(\hat{\theta}-\theta_o)$$

$$= nI^*(\tilde{\theta})(\hat{\theta}-\theta_o)$$

where $\tilde{\theta} = \tilde{\theta}(x^n)$ also lies on the line joining θ_o to $\hat{\theta}$. Since $\hat{\theta}$ is the M.L.E the first term in the expansion is 0. Using the resulting equation in the definition of Z_n gives

$$Z_n = \sqrt{n}(\theta^* - \theta) = \frac{1}{\sqrt{n}} I*(\tilde{\theta})^{-1} \nabla \log p(x^n \mid \theta_o)$$

$$= I*(\tilde{\theta})^{-1} S_n. \tag{6}$$

The transformation $I^*(\tilde{\theta})$ is the way we can convert our statements about Z_n whose structure we do not know into statements about S_n whose structure we do know: $S_n = (1/\sqrt{n}) \sum_{i=1}^{n} Y_i$ is a weighted sum of the i.i.d. random vectors $Y_i = \nabla \log p(X_i \mid \theta_o)$ and satisifes the moment condition

$$E\ S_n S_n^T = I(\theta_o),$$

because, for any i between 1 and n, $E\ Y_i Y_i^t = I(\theta_o)$, and $E\ Y_i = 0$. By equation (6), and the fact that inversion commutes with transposition, we have that

$$E\ Z_n I*(\theta^{**}) Z_n\ \chi_{\Omega_{\varepsilon,n}} = E\ S_n^T I*_n(\tilde{\theta})^{-1}\ I*(\theta^{*}*)\ I*(\tilde{\theta})^{-1}\ S_n\ \chi_{\Omega_{\varepsilon,n}}$$

$$= E\ S_n^T\ I^{-1}(\theta_o)\ S_n\ \chi_{\Omega_{\varepsilon,n}} + E\ S_n^T\ (A_n - I^{-1}(\theta_o))\ S_n\ \chi_{\Omega_{\varepsilon,n}} \tag{7}$$

where we define

$$A_n = I*(\tilde{\theta})^{-1}\ I*(\theta^{**})\ I*(\tilde{\theta})^{-1}.$$

Now, since $E\ S_n^T I^{-1}(\theta_o) S_n = d$, it remains to show that $E\ S_n^t I(\theta_o)^{-1} S_n \chi_{\Omega_\varepsilon^c}$ and $E\ S_n^T\ (A_n - I^{-1}(\theta_o))\ S_n\ \chi_{\Omega_{\varepsilon,n}}$ tend to zero as $n\ (\rightarrow \infty$. Note that $S_n^t I(\theta_o)^{-1} S_n$ is a positive quantity, and is uniformly integrable, since it converges in distribution and has expectation equal to the expectation of its limit, as in Chung [10] pp. 97. Then $E\ S_n^t I(\theta_o)^{-1} S_n \chi_{\Omega_\varepsilon^c} \rightarrow 0$, since $P(\Omega_\varepsilon^c \mid \theta_o) \rightarrow 0$.

It remains to show that the last term of (7) goes to zero as n increases.

Given the domain of integration this is easy. For, we have that

$$A_n \rightarrow I^{-1}(\theta_o)$$

in P_{θ_o} probability, and that the sequence $A_n \chi_{\Omega_{\varepsilon,n}}$ is bounded in norm. Also, we have that

$$S_n^T (A_n - I^{-1}(\theta_o))\ S_n\ \chi_{\Omega_{\varepsilon,n}} \rightarrow 0$$

in P_{θ_o} probability, by weak convergence to a constant and Slutsky's theorem. Since $S_n^t I^{-1}(\theta_o) S_n$ is uniformly integrable and $A_n \chi_{\Omega_{\varepsilon,n}}$ is bounded in norm, it follows that there exists a constant C so that

$$| S_n^T(A_n - I^{-1}(\theta_o))\, S_n \quad | \chi_{(*W_{\varepsilon,n}} \le C\ S_n^t I^{-1}(\theta_o) S_n ,$$

so

$$S_n^T(A_n - I^{-1}(\theta_o))\, S_n\ \chi_{\Omega_{\varepsilon,n}}$$

is uniformly integrable.

The second clause of the proposition follows from considering the integrand in (5): it converges in law to a chi-square with d degrees of freedom since the probability of the complement of the domain goes to zero. □

Appendix C

Here we give the statements and proofs of the lemmata which we use in the proofs of the propositions and the theorem. The first two are bounds on the rate of decrease of probabilities which occured in bounding (6) from above. It is assumed that $p(x \ \ | \theta)$ is continuous in θ for each fixed x.

Lemma C.1: (i) Assumptions 2 and 3 imply that, for any given $\delta > 0$,

$$P_{\theta_o}(\ | \hat{\theta} - \theta_o \ \ | > \delta) = O(1/n).$$

(ii) Also, assumptions 2 and 3 imply that there exists $\varepsilon > 0$ such that

$$P_{\theta_o}(p(X^n \ \ | \theta_o) \le e^{n\varepsilon} \sup_{\theta \in B_\delta^c} p(X^n \ \ | \theta)) = O(1/n).$$

(iii) Assumptions 4, and 7 imply that, for ε small enough,

$$P_{\theta_o}(\sup_{\theta \in B_\varepsilon} \ | \ | I^*(\theta) - I(\theta_o) \ | \ | > \delta) = O(1/n).$$

Remark : Conclusions (i) and (ii) are patterned after results of Wald [24], and Wolfowitz [27], respectively, so as to give rates of convergence.

Proof: First note that the event

$$\{ \ | \hat{\theta} - \theta_o \ \ | > \delta \ \}$$

is contained in the event

$$\{ \ p(X^n \ \ | \theta_o) \le \sup_{\theta \in B_\delta^c} p(X^n \ \ | \theta) \ \}.$$

So, to prove (i), it suffices to prove (ii). We use Chebyshev's inequality in a proof patterned after Wald. From assumptions

(1) and (2) select r sufficiently large that for some $\varepsilon', \eta' > 0$ we have

$$-\varepsilon' - E \sup_{||\theta - \theta_o|| > r} \log \frac{p(X \mid \theta)}{p(X \mid \theta_o)} > \eta'.$$

Cover the compact set

$$\{ \theta \mid ||\theta|| \le r, \ ||\theta - \theta_o|| \ge (*d \}$$

with finitely many small balls B_i centered at points θ_i, with radius ε_i, where i ranges from 1 to k, so that for some $\varepsilon'', \eta'' > 0$ we have

$$-\varepsilon'' - E \sup_{\theta \in B_i} \log \frac{p(X \mid \theta)}{p(X \mid \theta_o)} > \eta''.$$

From assumption (3), and the dominated convergence theorem, that can be done since the Kullback - Liebler number $E \log p(X \mid \theta_o)/p(X \mid \theta)$ is positive. If $0 < \varepsilon < \min(\varepsilon', \varepsilon'')$ and $\eta = \min(\eta', \eta'')$ then,

$$P_{\theta_o}(\sup_{|\theta - \theta_o| > \delta} \log \frac{p(X^n \mid \theta)}{p(X^n \mid \theta_o)} > -n\varepsilon)$$

$$\le P_{\theta_o}(\sup_{|\theta - \theta_o| > r} \log \frac{p(X^n \mid \theta)}{p(X^n \mid \theta_o)} > -n\varepsilon) + \sum_{i=1}^{k} P_{\theta_o}(\sup_{|\theta - \theta_i| < \varepsilon_i} \log \frac{p(X^n \mid \theta)}{p(X^n \mid \theta_o)} > -n\varepsilon)$$

$$\le P_{\theta_o}(\frac{1}{n} \sum_{i=1}^{n} \sup_{|\theta - \theta_o| > r} \log \frac{p(X_i \mid \theta)}{p(X_i \mid \theta_o)} - E \sup_{|\theta - \theta_o| > r} \log \frac{p(X \mid \theta)}{p(X \mid \theta_o)} > \eta)$$

$$+ \sum_{j=1}^{k} P_{\theta_o}(\frac{1}{n} \sum_{i=1}^{n} \sup_{|\theta - \theta_i| < \varepsilon_i} \log \frac{p(X_i \mid \theta)}{p(X_i \mid \theta_o)} - E \sup_{|\theta - \theta_i| < \varepsilon_i} \log \frac{p(X \mid \theta)}{p(X \mid \theta_o)} > \eta)$$

Since $\eta > 0$, Chebyshev's inequality may be applied to each term so as to upper bound the right hand side by O(1/n).

The proof of part (iii) is similar: Use the matrix norm which sums the absolute values of the entries and consider each $\sup_{\theta (m\!\delta B_\varepsilon} | i^*_{j,k}(\theta) - i_{j,k}(\theta_o) |$ separately. Let $\varepsilon > 0$ be less than any ξ which satisfies assumption 7. We modify the argument from (ii). Note that

$$P_{\theta_o}(\sup_{|\theta_o - \theta'| < \varepsilon} | i^*_{j,k}(\theta') - i_{j,k}(\theta_o) | > \eta)$$

$$\le P_{\theta_o}(\sup_{|\theta_o - \theta'| < \varepsilon} | i^*_{j,k}(\theta') - i^*_{j,k}(\theta_o) | > \frac{\eta}{2}) + P_{\theta_o}(| i^*_{j,k}(\theta_o) - i_{j,k}(\theta_o) | > \frac{\eta}{2})$$

The second term is O(1/n), by Chebyshev. For the first, choose ε so small that

$$E \sup_{|\theta_o - \theta'| < \varepsilon} | \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(x \mid \theta') - \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(x \mid \theta_o) | < \frac{\eta}{4}$$

then set up another application of Chebyshev's inequality. □

In most cases where we have wanted to we prove the convergence of a sequence of matrices we have used the matrix norm $||\cdot||$ on d x d matrices which sums the absolute values of the entries of the matrix. We denote the Euclidean norm on the parameter space with the same symbol, since the argument will indicate which meaning is intended. Clearly, by the equivalence of norms on Euclidean spaces, lemma 1 remains true under any choice of Euclidean norm on the parameter space or on the space of $d \times d$ matrices. In particular, to prove the next lemma it is more convenient to use either the norm on matrices which takes the supremum of the entries, or the norm which takes the largest of the absolute values of the eigenvalues.

In the next lemma we state three routine results which can be used to prove the two propositions used in the theorem. Parts (i) and (ii) amount to an equivalence of neighbourhood bases about the true value of the parameter.

Specifically, let

$$B_{\rho'} = \{\theta \in R^d \mid (\theta-\theta_o)^T I(\theta_o)(\theta-\theta_o) < \rho'\}$$

$$B^*_{\rho} = \{\theta \in R^d \mid (\theta-\hat{\theta})^T I^*(\hat{\theta})(\theta-\hat{\theta}) < \rho\}.$$

We have assumed that $I(\theta_o)$ is positive definite. This means $B_{\rho'}$ and B^*_{ρ} are neighbourhoods of θ^* and θ_o and all eigenvalues of $I(\theta_o)$ are positive.

The third part of the lemma sandwiches an approximation of $I(\theta_o)$ between two other approximations which are evaluated at the M.P.L.E. and weighted by factors close to 1. The symbol < used between matrices means that the bilinear form induced by one matrix is greater than the biliniear form induced by the other matrix.

Lemma C.2: Assumptions 1,2,3,4,6, and 7 imply

(i): Given $\rho > 0$ there exists $\rho' > 0$ such that

$$B^*_{\rho} \supseteq B_{\rho'}$$

with probability at least $1 - c/n$ for all large n.

(ii) : Given $\rho' > 0$ there exists a $\rho > 0$ such that

$$B^*_{\rho} \subseteq B_{\rho'}$$

with probability at least $1 - c/n$ for all large n.

(iii) : Given $\tau \in (0,1)$ we have that

$$(1+\tau)I^*(\hat{\theta}) \geq I^*(\theta^{**}) \geq (1-\tau)I^*(\hat{\theta}),$$

with probability greater than $1 - c/n$ for all large n, where θ^{**} is on the line joining θ_o to $\hat{\theta}$.

Proof: Parts (i) and (ii) follow from routine calculations with inner products on $\mathbf{R}^d$, using lemma C.1 parts (i) and (ii).

Part (iii) follows from considering disjoint open sets about $I(\theta_o)$, $(1+\tau)I(\theta_o)$, and $(1-\tau)I(\theta_o)$ and choosing n so large that the estimates of them lie in the open sets. □

References

[1] J. Aitchison, ''Goodness of prediction fit,'' *Biometrika* vol. 62- 3, pp. 547-554, December 1975.

[2] R. R. Bahadur, ''Some limit theorems in statistics,'' in *Regional Conference Series in Applied Mathematics* , Philadelphia: Society for Industrial and Applied Mathematics, 1971.

[3] A. R. Barron,''Are Bayes Rules Consistent in Information?'' in *Problems in Communications and Computation* , T. M. Cover and B. Gopinath, Ed.
New York: Springer - Verlag, 1987, pp. 85-91.

[4] A. R. Barron, ''The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions,'' submitted to *Ann. Stat.* 1988.

[5] A. R. Barron and T. M. Cover, ''A bound on the financial value of information, .IP [6] 5 R. E. Blahut, *Principles and Practice of Information Theory.* Reading:
Addison - Wesley, 1987.

[7] N. N. Ĉencov, *Statistical Decision Rules and Optimal Inference.* Providence: American Mathematical Society, 1981.

[8] H. Chernoff, ''On the distribution of the likelihood ratio,'' *Ann. Math. Statist.* vol. 25-3, pp. 573-578, September 1954.

[9] H. Chernoff, ''Large sample theory: parametric case,'' *Ann. Math. Statist.* , vol. 27-1, pp. 1-22, March 1956.

[10] K. L. Chung, *A Course in Probability Theory.* New York: Accademic Pppress, 1974.

[11] H. Cramer, *Mathematical Methods of Statistics.* Princeton: Princeton, 1946.

[12] I. Csiszar, ''Information-type measures of difference of probability distributions and individual observations,'' *Studia Sci. Math. Hungar.*, vol. 2 , pp. 299-318, 1967.

[13] L. D. Davisson, ''Universal noiseless coding,'' *IEEE Trans. Inform. Theory,* vol. IT-19, pp. 783-795, November 1973.

[14] N. G. De Bruijn, *Asymptotic Methods in Analysis.* New York: Dover, 1958.

[15] J. Kelly, ''New interpretation of information rate,'' *Bell Syst. Tech. Journal,* vol. 35, pp. 917-926, July 1956.

[16] R. E. Krichevsky and V. K. Trofimov, ''The performance of universal encoding,''*IEEE Trans. Inform. Theory,* vol. IT-27, pp. 199-20 7, March 1981.

[17] S. Kullback, *Information Theory and Statistics.* New York: Wiley,
1959.

[18] S. Kullback, J. C. Keegel, J. H. Kullback, *Topics in Statistical Information Theory.* Berlin: Springer-Verlag, 1980.

[19] E. L. Lehmann, *Theory of Point Estimation.* New York: Wiley, 19 83.

[20] J. Rissanen, "Universal coding, information, prediction, and estimation, 84.

[21] S. M. Stigler, "Laplace's 1774 memoir on inverse probability," *Statistical Science* , vol. 1-3, pp. 359-378, August 1986.

[22] Tierney and Kadane, " Accurate approximations for posterior moments and marginal densities," *JASA* vol. 81 pp. 82-86, March 1986.

[23] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Amer. Math. Soc.*
vol. 54 pp. 426-482, November 1943.

[24] A. Wald, "Note on the consistency of the maximum likelihood estimate,'

[25] A. M. Walker, "On the asymptotic behaviour of posterior distributions, .IP [26] S. S. Wilks, *Mathematical Statistics.* New York: Wiley, 1962.

[27] J. Wolfowitz, "On Wald's proof of the consistency of the maximum likelihood estimate," *Ann. Math. Statist.*, vol. 20-4 , pp. 601-602, December 1949.