Chapter 2

Optimal Estimating Equations

Estimating functions, defined by Godambe and Kale (1991) as any real function g of both the observations $\mathbf{X}$ and the parameter θ, $g = g(\theta, \mathbf{X})$, provide a unifying structure connecting some of the classical estimation methods of statistical inference. These include maximum likelihood, least squares and method of moments. The following formulations illustrate these estimators as the roots of their respective estimating equations.

2.1 Estimating functions

Maximum Likelihood

In a single parameter model the maximum likelihood estimator $\hat{\theta}(\mathbf{x})$ is defined as

$$\hat{\theta}(\mathbf{x}) = \arg\ \max_{\theta \in \boldsymbol{\Theta}} \mathrm{L}(\theta|\mathbf{x})$$

which under regularity can be found as the solution to the *score equation*

$$\frac{\partial}{\partial \theta} l(\theta|\mathbf{x}) = 0,$$

where $l(\theta|\mathbf{x})$ is the log-likelihood.

Least Squares

Given a model, $E[\mathbf{Y}] = \mathbf{X}\theta$, the least squares estimator is defined as the value of θ which minimizes

$$(\mathbf{Y} - \mathbf{X}\theta)'(\mathbf{Y} - \mathbf{X}\theta).$$

After differentiation the normal equations are obtained,

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\theta = 0,$$

the solution of which is the least squares estimator of θ.

Method of Moments

Let $x_1 \ldots, x_n$ be an *i.i.d.* sample from a distribution with parameter θ. If the first q moments of the distribution $\alpha_i(\theta_1, \ldots \theta_q), i = 1, \ldots, q$ exist then the method of moments estimator is obtained by solving the set of equations

$$\alpha_i(\theta_1, \ldots, \theta_q) - a_i = 0, \quad \text{for } i = 1, \ldots, q,$$

where a_i are the corresponding observed sample moments.

In all of the above examples the estimator of θ is obtained by setting the estimating equations equal to zero and solving for θ.

Estimation is usually approached by stating a set of optimality criteria:- sufficiency, unbiasedness, minimum variance to name but a few -, deriving an estimator, and then comparing the properties of the estimator to the optimality criteria. Godambe and Kale (1991) point out that most estimation methods are essentially *ad hoc*, in that the selection of a method does not follow from the optimality criteria. Given a new problem there is no theory to say which estimation procedure should be used.

An advantage of approaching problems through estimating functions is that optimality properties can be defined in terms of the estimating functions themselves.

Another justification is that major parts in the proofs of the consistency and asymptotic normality of the *m.l.e.* involve facts about the score function (possibly the most important estimating function). In particular in an *i.i.d.* sample set up, the score function is an *i.i.d.* sum, so that the Law of Large Numbers, and the Central Limit Theorem will hold quite generally. These important properties of the score function argue that it in particular, and consequently estimating functions in general are worthy of study in their own right (Small & McLeish, 1988).

2.2 A hierarchy of inferential frameworks

It will be a convention in the thesis to write the model parameter θ as $\theta = (\psi, \lambda)$ where ψ is the parameter of interest and λ is the nuisance parameter. The phrase "at the model" will mean the true but unknown value of θ, denoted by $\theta_0 = (\psi_0, \lambda_0)$.

A parametric inferential framework will be assumed. To be precise, there is a probability space $(\Omega, \mathcal{F}, P_\theta)$, where P_θ is a probability measure $P_\theta \in \mathcal{P}$ defined on the σ-field $\mathcal{F}$. $\mathcal{P}$ is indexed by the set Θ, and there is assumed to be a one to one correspondence between the elements of $\mathcal{P}$ and the elements of Θ, consequently there is no ambiguity in denoting an element of $\mathcal{P}$ as P_θ. Θ will be assumed to be the Cartesian product of open sets in $\mathbb{R}$. Now assume that there is a probability measure $P_{\theta_0} \in \mathcal{P}$, on $\mathcal{F}$ satisfying the conditions

- P_{θ_0} is σ-finite.
- $P_\theta \in \mathcal{P}$ is absolutely continuous with respect to P_{θ_0} for all θ.
- The ratios $\frac{dP_\theta}{dP_{\theta_0}}$ are square integrable with respect to P_{θ_0} for all θ.

The above mild conditions ensure firstly the existence and uniqueness (up to sets of P_{θ_0}-measure 0) of the likelihood ratios with respect to P_{θ_0}, and secondly that the likelihood ratios are elements of the space $\mathcal{L}_2(P_{\theta_0})$. The space $\mathcal{L}_2(P_{\theta_0})$ is defined as the space of all real valued functions f such that $\int_\Omega f^2 dP_{\theta_0} < \infty$.

Usually densities will be defined with respect to Lebesgue measure, so that the density $dP_\theta = p_\theta$ satisfies $\int_\Omega p_\theta dx = 1$.

The above assumptions will be supplemented with others, the most notable one being that the densities p_θ will be assumed to satisfy Cramér-type regularity conditions. In particular it will always be assumed that if $g(\theta, \mathbf{x})$ is any estimating function, then the order of differentiation and integration may be changed in the following expectation, so that

$$
\begin{aligned}
E_\theta\left[g'(\theta)\right] &= \int_\Omega g'(\theta) + \left(g(\theta)\frac{p'_\theta}{p_\theta} - g(\theta)\frac{p'_\theta}{p_\theta}\right) p_\theta dx \\
&= \int_\Omega \left(g(\theta)p_\theta\right)' dx - \int_\Omega \left(g(\theta)\frac{p'_\theta}{p_\theta}\right) p_\theta dx \\
&= \frac{\partial}{\partial\theta}\int_\Omega g(\theta)p_\theta dx - \int_\Omega g(\theta)U(\theta)p_\theta dx \\
&= 0 - E[g\,U] = -E[g\,U], \qquad (2.1)
\end{aligned}
$$

where $U(\theta) = p'(\theta)/p(\theta)$ is the score function for the model, that is the derivative of the log-likelihood. The preceding regularity conditions will be referred to collectively as the *basic regularity conditions.*

In order to derive estimative strategies in problems that are not amenable to reduction by sufficiency, a series of results will be outlined that are suggestive of a general methodology. For the sake of simplicity these results will be stated for the case where the parameter of interest and the nuisance parameter are both univariate, although they extend to multivariate parameters without difficulty.

A *Type 0* model is one in which there are no nuisance parameters. Godambe (1960) provided the first fundamental optimality result for estimating equations in this model. He restricted attention to the class of *unbiased* estimating functions $\mathcal{G}_0$ where an estimating function is an element of $\mathcal{G}_0$ if it satisfies the condition:

$$E_\theta\left[g(\theta, \mathbf{X})\right] = 0 \quad \text{for all } \theta \in \Theta \tag{2.2}$$

The reason for restricting attention to this particular class of functions is that they will provide consistent estimates. Recall that consistency of the root of the estimating equation is dependent on the fact that the estimating equation is an unbiased estimate of 0, see (1.3).

Having restricted attention to $\mathcal{G}_0$, Godambe suggested an intuitive optimality criteria as a means of discriminating between the elements of $\mathcal{G}_0$. The optimality goal is to find the function $g^* \in \mathcal{G}_0$ which minimizes the quantity

$$\frac{E_\theta[g^2]}{\left(E_\theta[\partial g/\partial\theta]\right)^2} \tag{2.3}$$

over all $g \in \mathcal{G}_0$ and $\theta \in \Theta$. It is reasonable to describe the criteria as "intuitive" since it is the asymptotic variance of the root of the estimating equation. Hence the minimization of Godambe's criteria leads to an estimator with minimum asymptotic variance within the class of unbiased estimating functions. Various other justifications of the criteria have been given (Small & McLeish, 1988; Godambe & Heyde, 1987; Godambe & Kale, 1991). The inverse of (2.3) will be defined as the *G-information* in g, and written I_g. Consequently maximizing the G-information is equivalent to minimizing Godambe's criteria (2.3). In this thesis the viewpoint will be to maximize informations rather than minimize variances, even though they are of course equivalent.

The key result established by Godambe is that under mild regularity conditions, in the one parameter model $p(x, \theta)$, the score function is the optimal estimating equation in $\mathcal{G}_0$. That is, writing the score $\frac{\partial}{\partial\theta}log(p)$ as U

$$\frac{E_\theta[U^2]}{\left(E_\theta\left(\frac{\partial}{\partial\theta}U|\theta\right)\right)^2} \leq \frac{E_\theta[g^2]}{\left(E_\theta\left(\frac{\partial}{\partial\theta}g|\theta\right)\right)^2} \quad \text{for all } g \in \mathcal{G}_0 \text{ and } \theta \in \Theta \tag{2.4}$$

The proof is simple: (Compare the following proof to a careful one of the optimality of the **root** of the score equation, in other words the optimality of the *m.l.e.*). Define the *relative G-information* in a function g as I_g/I_U, then I_g/I_U equals $\mathrm{corr}^2(\mathrm{U},\mathrm{g})$. This follows from the assumed regularity, (2.1), which implies that for any unbiased function g, $E_\theta(\frac{\partial}{\partial\theta}g) = -E[g\,U]$. Thus the relative G-information can never be greater than 1, with equality if and only if g is a linear function of U.

The proof demonstrates one of the great advantages of taking an estimation equation viewpoint for inference in that it provides "optimality without tears".

Godambe also linked his optimality criteria to the Cramér-Rao inequality: for any function of the data alone, say $h(\mathbf{x})$, define $g(\mathbf{x},\theta) = h(\mathbf{x}) - E_\theta[h(\mathbf{X})]$. Then substituting g into the right hand side of (2.4) gives,

$$\frac{E[U^2]}{(E[U']^2)} \leq \frac{\mathrm{Var}[\mathrm{h}(\mathbf{X})]}{\left(\frac{\partial}{\partial\theta}\mathrm{E}[\mathrm{h}(\mathbf{X})]\right)^2}. \tag{2.5}$$

By again using the score identity $E[U^2] = -E[U']$, and rearranging (2.5), one obtains the Cramér-Rao inequality

$$\frac{\left(\frac{\partial}{\partial\theta}E[h(\mathbf{X})]\right)^2}{E[U^2]} \leq \mathrm{Var}[\mathrm{h}(\mathbf{X})]. \tag{2.6}$$

A *Type I* nuisance parameter problem is defined to be one in which, treating ψ fixed, there exists a complete sufficient statistic $T(\mathbf{X})$ for the nuisance parameter λ, for all values of ψ, where T is a function of the data alone. This assumption asserts that the density factors as

$$p(\mathbf{x},\theta) = f_{\mathbf{X}|T}(\mathbf{x};t,\psi)\,h(t;\psi,\lambda)$$

and that the family of densities of t is complete in λ for each fixed ψ.

Within the class of functions $\mathcal{G}'_\infty$, where $\mathcal{G}'_\infty$ consists of all functions g that depend on the data and parameter of interest alone, and that are unbiased in the sense that

$$E_\theta[g(\psi, \mathbf{x})] = 0 \text{ for all } \theta \in \Theta, \tag{2.7}$$

Godambe (1976) showed that the conditional score $U_c(\psi, t, \mathbf{X}) = \frac{\partial}{\partial \psi} \, ln \, f_{\mathbf{X}|T}(\mathbf{x}; t, \psi)$ is the optimal estimating equation according to the optimality criteria (2.3).

As an example, consider a strata in the Neyman-Scott (Example 2) problem. Then $\overline{x}_i$ is a complete sufficient statistic for the nuisance parameter λ_i and it is a simple task to establish that the joint density of the x_{ij} factors into

$$\overbrace{\left[\frac{1}{2\pi\psi}\right]^{\frac{n-1}{2}} \frac{1}{\sqrt{n}} \exp\left(\frac{-1}{2\psi}\left[(n-1)s_i^2\right]\right)}^{A} \overbrace{\frac{\sqrt{n}}{\sqrt{2\pi\psi}} \exp\left(-\frac{n}{2\psi}\left(\overline{x}_i - \lambda_i\right)^2\right)}^{B}$$

where $(n-1)s_i^2 = \sum (x_{ij} - \overline{x}_i)^2$. The function under the second brace is the density of $\overline{x}_i$ so that the conditional score function is

$$U_c(\psi) = \frac{\partial}{\partial \psi} \log\left(\left[\frac{1}{2\pi\psi}\right]^{\frac{n-1}{2}} \frac{1}{\sqrt{n}} \exp\left(\frac{-1}{2\psi}\left[(n-1)s_i^2\right]\right)\right),$$

which simplifies to

$$-\frac{n-1}{2}\frac{1}{\psi} + \frac{1}{2\psi^2}(n-1)s_i^2.$$

Note that λ has been eliminated, and the solution to the equation is $\hat{\psi} = s_i^2$.

In many models the complete sufficient statistic for the nuisance parameters will depend on the parameter of interest, so that $T = T(\psi)$; these are the *Type II* problems. As an example of this situation consider two independent observations (X, Y) from exponential distributions with means $\frac{1}{\psi\lambda}$ and $\frac{1}{\lambda}$ respectively. A complete sufficient statistic for λ holding ψ fixed is $\psi X + Y$, and it can be shown that the conditional score is $U_c(\psi,\lambda) = -\lambda\frac{\psi X - Y}{2\psi}$. It is important to note that the conditional score depends on the nuisance parameter.

Lindsay (1982) established the following results for this class of problems. He showed that the conditional score, defined as

$$U_c(\psi,\lambda) = U_0(\psi,\lambda) - E_\psi(U_0(\psi \mid T(\psi)) \tag{2.8}$$

(where U_0 is the score for ψ) is unbiased in the sense that

$$E_\theta[U_c(\psi,\lambda_0)] = 0 \text{ for all } \theta \text{ and } \lambda_0.$$

It is optimal according to the information criteria (2.3) within the class of functions $\mathcal{G}_\infty$ whose elements g, satisfy $E_\theta[g(\psi,\lambda_0)] = 0$ for all θ and λ_0. Why restrict attention to the class $\mathcal{G}_\infty$ of functions rather than the wider class, $\mathcal{G}_0$, of unbiased functions satisfying $E_\theta[g(\psi,\lambda)] = 0$ for all θ? The first reason is that in nuisance parameter problems, where the nuisance parameters must be estimated, the substitution version of an unbiased estimating function will not in general be unbiased. Consequently estimating functions in $\mathcal{G}_0$ will potentially have inconsistent roots. Reasons as to why $\mathcal{G}_\infty$ is the right class of functions to use in nuisance parameter problems include

- The class $\mathcal{G}_\infty$, treating ψ as fixed, satisfies Lehmann's first order ancillarity condition (1981).
- If there is an estimate of λ, say $\tilde{\lambda}_\psi$, that is independent of $g(\psi,\lambda;\mathbf{x})$ then

  $$E_\lambda[g(\tilde{\lambda})] = E\left[E_\lambda[g(\tilde{\lambda} \mid \tilde{\lambda})]\right] = 0 \text{ for all } \lambda$$

  so that the substitution version of g retains this strong unbiasedness property.
- It is precisely the E-ancillary class referred to in 1.18

- Admittedly, this is somewhat tautological, but using $\mathcal{G}_\infty$ gives the conditional score as the optimal estimating equation. The circularity of the argument can be alleviated by the recognition that conditionality as a principle arises in a variety of areas of statistical inquiry. In particular, the theory of locally most powerful testing gives rise in certain situations to conditional score functions as optimal test statistics. (There are also strong philosophical grounds for the conditionality principle).

Since the optimal estimating equation now depends on the unknown nuisance parameter λ it can no longer be globally optimal. This means that the conditional score function evaluated at λ_0, $U_c(\psi, \lambda_0)$, is still unbiased for all θ, but typically optimal only at $\theta = (\psi, \lambda_0)$. In this case a locally optimal estimating equation for ψ can only be constructed from U_c in the Type II problem if the true value of λ_0 is known. One of Lindsay's (see (Lindsay, 1982, p.505) for the full set of assumptions) results was that if an estimate for λ, say $\tilde{\lambda}$ that is a function of the data only through $T(\psi)$ (which is certainly the case for the *m.l.e.* of λ, $\hat{\lambda}_\psi$) is substituted into U_c then the resulting estimating equation is still unbiased. Consequently it will provide a consistent root. This approach will be referred to as the "substitution strategy", and mimics exactly what is done in standard maximum likelihood estimation.

Remark: The substitution version of U_c, that is $U_c(\tilde{\lambda})$, is not locally optimal. The difference between $U_c(\tilde{\lambda})$ and $U_c(\lambda_0)$ is a measure of the price that has been paid as a result of ignorance about the true value of λ, and it is of interest to investigate this price. The critical difference between Type I and Type II problems is the loss of global optimality of the score equation because of its dependence on the nuisance parameters. This indicates that in more general problems, if local optimality is to be achieved, then estimating functions that depend on both the nuisance parameter and parameter of interest must be considered.

Strictly speaking U_c in the Type II problem is not an estimating equation due to its dependence on λ and such a function will be referred to as an *estimative kernel* .

The previous results are conveniently summarized in Table 2.1.

Since the conditional score evaluated at λ_0 is the locally optimal estimating equation in the Type II model, this suggests that in more general problems one should search for an estimative kernel that mimics the properties of U_c. Also, any theory in a general setting can at most hope to be based around local optimality.

In order to examine the theory of estimation equations beyond Type II models some groundwork must be laid. Given the fact that only local optimality, at most, will be achievable an overview of the theory of locally minimum variance unbiased estimation is required. Results from four papers will be discussed. These are Barankin (1949), Stein (1950), Rao (1952) and Bahadur (1957). The results vary in their generality, but there is considerable overlap. All the results are stated in a univariate formulation, that is, one in which there is a single parameter θ. Furthermore, attention is confined to functions which are to be estimated by statistics, where a statistic is allowed to depend only on the data, and not on the parameters. Even though the results were not explicitly stated in terms of estimating functions, they are in fact applicable in this setting.

The inferential framework common to all the papers is the existence of a sample space Ω, and a sigma field $\mathcal{F}$ of subsets of Ω. There is a family of densities, say $\{p_\theta\}$ with respect to a σ-finite measure μ defined on $\mathcal{F}$, the densities being indexed by a parameter $\theta \in \Theta$.

Table 2.1: Optimal estimating functions within the inferential hierarchy.

Model	Parameters	Estimating Functions	Unbiasedness Condition	Optimal Equation	Type of Optimality
Type 0	θ	$g(\theta, \mathbf{x})$	$E_\theta[g(\theta, \mathbf{x})] = 0 \; \forall \; \theta$	$U(\theta, \mathbf{x})$	Global
Type I	$\theta = (\psi, \lambda)$	$g(\psi, \mathbf{x})$	$E_\theta[g(\psi, \mathbf{x})] = 0 \; \forall \; \theta$	$U_c(\psi, \mathbf{x})$	Global
Type II	$\theta = (\psi, \lambda)$	$g(\psi, \lambda, \mathbf{x})$	$E_\theta[g(\psi, \lambda_0, \mathbf{x})] = 0 \; \forall \; \theta, \lambda_0$	$U_c(\psi, \lambda_0, \mathbf{x})$	Local

Define a real valued measurable (μ) function f_0 on Ω to be *unbiased* for the function g on Θ if it satisfies

$$\int_\Omega f_0\, p(\theta) d\mu = g(\theta) \text{ for all } \theta \in \Theta$$

The objective is to establish the existence and characterization of an unbiased estimate, say f^* that is in some sense "best" at θ_0. The optimality condition is usually taken to be

$$\int_\Omega (f^* - g(\theta_0))^2 p_{\theta_0} d\mu \leq \int_\Omega (f - g(\theta_0))^2 p_{\theta_0} d\mu$$

for any other unbiased estimate f. If this best estimate f^* does not depend on θ_0 then it is said to be a *Uniformly Minimum Variance Unbiased Estimator* of $g(\theta)$ (UMVUE). If the best estimate depends on θ_0, as will usually be the case, then it is termed a *Locally Minimum Variance Unbiased Estimator* (LMVUE).

Barankin considers a wider definition of bestness. In particular he seeks to find a function f^* that minimizes

$$\int_\Omega |f^* - g(\theta_0)|^s p_{\theta_0} d\mu \leq \int_\Omega |f - g(\theta_0)|^s p_{\theta_0} d\mu$$

where s is a fixed number $s > 1$.

He assumes that

- The likelihood ratios $\pi_\theta(x) = \frac{p_\theta(x)}{p_{\theta_0}(x)}$ are defined almost everywhere μ and for all θ.

- $\pi_\theta(x) \in \mathcal{L}_r$.

Here, $\mathcal{L}_r$ is the set of functions that have integrable absolute r^{th} powers with respect to the measure ν defined by $\nu(A) = \int_A p_{\theta_0} d\mu$, where $r = \frac{s}{s-1}$.

Given any function $g(\theta)$ he presents *necessary and sufficient* conditions for the existence of an unbiased estimate. These conditions are equivalent to those of

Banach (1987, p.34, Theorem 4) for the existence of a certain linear functional G_0. This functional, when applied to the ratios π_θ yields $g(\theta)$, and can be expressed according to the Riesz representation theorem as $G_0(\zeta) = \int_\Omega f^* \zeta d\nu$, where f^* is the LMVUE. A lower bound for the norm of any unbiased estimate of $g(\theta)$ is derived. Given the existence of at least one unbiased estimate, he then establishes the existence of a unique unbiased estimator that has norm equal to the lower bound, and is thus best at θ_0.

Assuming the existence of the derivatives of the π_θ and their membership of the class $\mathcal{L}_r$ (for additional regularity conditions see (Barankin, 1949, Theorem 5)) Barankin gives an explicit construction of the lower bound for the norm of unbiased estimates. This provides a generalization of the classical Cramér-Rao lower bound for the variance of unbiased estimates. Furthermore, assuming the existence of second derivatives of π_θ he then generalizes and constructs the first of the Bhattacharyya (1946) lower bounds for the variance of unbiased estimates of $g(\theta)$.

An explicit construction is presented in his Theorem 7 for the best estimate at θ_0, that is used in Section 8 of the paper to provide what is perhaps the most important result for use in this thesis. The case $s = 2$ is considered, so that the minimization criterion is the familiar one of variance minimization. Denote by Λ_∞ the closed linear manifold in $\mathcal{L}_2$ spanned by the π_θ, so that Λ_∞ is the smallest closed linear subspace that contains $\pi_\theta, \theta \in \Theta$. Let $\mathcal{M}_2 \subseteq \mathcal{L}_2$ denote the class of unbiased estimators of $g(\theta)$. The following theorem presents a characterization of the LMVUE of $g(\theta)$.

Proposition 2.1 *If $\mathcal{M}_2$ is non-empty, then the* LMVUE *of $g(\theta)$ is the unique element of $\mathcal{M}_2$ that lies in Λ_∞.*

It is an immediate consequence that if g is any function lying in Λ_∞ then it is the unique LMVUE of its own expectation.

The paper by Stein considers the estimation problem only for the case $s = 2$. Rather than use the Hahn-Banach theorem to establish the existence of the fundamental linear functional, he states his theorem in terms of the existence of the LMVUE if and only if the appropriate linear functional exists. Stein also presents two corollaries that essentially state that if the function to be estimated $g(\theta)$ is a linear combination or difference of terms that are expressible as $\sum_i a_i E_{\theta_i}[\pi(\theta)]$ then these are best estimated by $\sum_i a_i \pi(\theta)$, and that this will hold true as limiting arguments are applied to the θ_i. In particular if, for each θ_1

$$g(\theta_1) = \sum_{i=0}^{m} a_i \left[\frac{\partial^i}{\partial \theta^i} E_{\theta_1}[\pi_\theta] \right]_{\theta=\theta_0},$$

then the LMVUE of $g(\theta)$ is

$$\sum_{i=0}^{m} a_i \left[\frac{\partial^i}{\partial \theta^i} \pi_\theta \right]_{\theta=\theta_0}.$$

The Bahadur paper (1957) is concerned with *uniform* minimum variance unbiased estimation, though the case of locally minimum variance unbiased estimation is addressed in the first half of the paper. A degree of simplicity is achieved through the assumption of the existence of at least one unbiased estimate of the function under consideration, so that the proofs are not concerned with existence, but rather the characterization of unbiased estimates. Bahadur demonstrates that the LMVUE of $g(\theta)$ is the projection of any unbiased estimate of $g(\theta)$ onto the space Λ_∞, and furthermore that the LMVUE is unique.

Two pertinent comments consider the result of projecting an unbiased estimate, firstly onto a space that contains Λ_∞, and secondly onto a subspace contained in Λ_∞. For the first situation, if $\Lambda'_\infty \supseteq \Lambda_\infty$ and the projection onto Λ'_∞ is denoted by Π', then if $g(x)$ is an unbiased estimate of $h(\theta)$ then so is $\Pi' g$, and furthermore $||\Pi' g|| \leq ||g||$, with strict inequality unless $\Pi' g = g$. The idea is that projection onto a subspace that contains Λ_∞ ensures the reduction in variance of the original estimate, and also that the projected version of the estimator retains the unbiasedness properties of its originator. This idea constitutes an extension of the Rao-Blackwellization of a statistic, where a statistic is projected onto the space of T measurable functions, where T is a sufficient statistic for the parameter θ.

For the second case, in which the space that the estimator is projected onto, say Λ''_∞, is contained in Λ_∞, then the result is that there exists a function $g^* \in \Lambda''_\infty$, such that $\Pi'' g = g^*$ for any unbiased g. Furthermore, $||g^*|| \leq ||g||$ with strict equality only for $g = g^*$. Projection onto Λ''_∞ will provide a lower bound to the variance of the LMVUE. Consider now a sequence of orthogonal elements V_i, where for each $i, V_i \in \Lambda_\infty$. The V_i generate an increasing sequence of subspaces, say $\Lambda_i : i = 1, 2 \cdots = \langle V_1, V_2 \cdots V_i \rangle$, which will provide a decreasing sequence of lower bounds. It is of interest to determine if these lower bounds converge to the variance of the LMVUE. This will be the case if and only if each of the likelihood ratios π_θ is contained in $\bigvee_i \Lambda_i$, where $\bigvee_i \Lambda_i$ is the smallest subspace containing $\{V_i : i = 1, 2 \cdots\}$.

The paper by Rao (1952) contains a very useful characterization of LMVUE. Assume there is a specified function of θ, say $g(\theta)$ that is of interest. Denote the class of all unbiased estimators of $g(\theta)$ by $\mathcal{M}_2$. Let $\mathcal{N}$ denote the class of all functions with expectation 0. That is $g_0 \in \mathcal{N}$ if and only if $E_\theta[f] = 0$ for all θ. Then the result is that in order for an estimator $T \in \mathcal{M}_2$ to be a LMVUE of $g(\theta)$ with minimum variance at θ_0 it is necessary and sufficient that $\mathrm{Cov}_{\theta_0}(T, f) = 0$ for all $f \in \mathcal{N}$, provided all the functions are square integrable. Furthermore the LMVUE is essentially unique.

This provides an alternative characterization of LMVUE to Barankin's. Recall that Barankin identifies them as the elements of the space generated by the likelihood ratios, whereas Rao characterizes them as the functions that are orthogonal to the elements of $\mathcal{N}$. Note that the class $\mathcal{N}$ are precisely those functions that are first order ancillary in the sense of Lehmann (1981).

2.3 Multivariate estimating equation generalizations

For the multivariate case we have $\theta = (\psi, \lambda)$, where ψ and λ will be taken to be r and s dimensional respectively. The theory to be developed is a local theory; it will assume that the value of the nuisance parameter is known, and equal to λ_0. The theory will use the optimal properties of the conditional score in the Type II problems as a guide to the darkness. One immediate question of interest is under what circumstances will the local optimality properties not depend on having assumed that λ_0 was known, and so consequently become global or uniform optimality properties in λ. Since λ_0 is never really known, and must be estimated, it is critical to asses the sensitivity of the locally optimal estimating equations to the replacement of λ_0 with an estimate $\tilde{\lambda}$. This issue will be addressed in Chapter 4, where a series of results will be presented which demonstrate the sense in which the local estimating equations are insensitive.

The Hilbert space of square integrable functions (dP_{λ_0}) will be denoted by $\mathcal{L}_2$, where the inner product in this space is defined by

$$< g_1, g_2 >= \int_\Omega g_1(\psi, \lambda_0) g_2(\psi, \lambda_0) p(\psi, \lambda_0) dx.$$

Since the consistency of the root of an estimating equation is intimately connected to its unbiasedness, attention will be restricted to the class of mean 0 estimating functions. More precisely, an estimating equation $g \in \mathcal{L}_2$ will be said to be unbiased or an element of $\mathcal{G}_0$ if

$$E_{\psi\lambda_0}[g(\psi, \lambda_0)] = 0 \text{ for all } \psi \text{ and } \lambda_0.$$

In general $\mathcal{G}_0$ need not be a closed space, that is it may not contain its limit points, but this is a fundamental requirement for any theory that utilizes projections. In order to guard against this possibility $\mathcal{G}_0$ will be henceforth be redefined as the closure of $\mathcal{G}_0$, that is $\overline{\mathcal{G}_0}$. The purpose of this is to ensure that $\mathcal{G}_0$ is a Hilbert space, a space which is by definition closed.

Since ψ is of dimension r we consider vector valued unbiased estimating functions lying in the r-fold Cartesian product of the space space $\mathcal{G}_0$, and will denote the space of all such functions by $\mathcal{G}_0^r$. The Cartesian product can be written as a direct sum, and the direct sum of a family of Hilbert spaces is well known to be itself a Hilbert space (Halmos, 1951, §9). It follows that $\mathcal{G}_0^r$ is a Hilbert space. For any two functions g_1 and $g_2 \in \mathcal{G}_0^r$ the inner product is given by $\langle g_1, g_2\rangle = \sum_{i=1}^r \langle g_{1i}, g_{2i}\rangle$, where the components g_{ai}, $a = 1, 2$, are all in $\mathcal{G}_0$. An equivalent way of the writing the inner product in matrix notation is $\langle g_1, g_2\rangle = \mathrm{tr}E_{\lambda_0}[g_1\, g_2^T]$. It follows that for an unbiased estimating function to lie in $\mathcal{G}_0^r$ the condition

$$\mathrm{tr}E_{\psi\lambda_0}[g(\psi, \lambda_0)g(\psi, \lambda_0)^T] < \infty$$

must be satisfied.

Since the objects of attention are estimating functions, that are by definition unbiased, that is for any g, $E[g] = 0$ it is natural to define, for ψ fixed, the space Λ_∞ as the closed linear subspace generated by the *centered likelihood ratios*

$$\left\{ \frac{p(\psi, \lambda, x)}{p(\psi, \lambda_0, x)} - 1 \right\}.$$

This definition ensures that any function $g \in \Lambda_\infty$ is itself unbiased. Furthermore, if the centered likelihood ratios are augmented with a constant vector, say 1, then the space $\Lambda_\infty + 1$ (here + is being used for vector sum) is identical to the space generated by the plain likelihood ratios $\frac{p(\psi,\lambda,x)}{p(\psi,\lambda_0,x)}$. This means that projecting an unbiased estimating equation onto the spaces Λ_∞ and $\Lambda_\infty + 1$ is equivalent, so that the theory of locally minimum variance unbiased estimation, established by projection onto the the space generated by the likelihood ratios, will be will be repeated if unbiased estimating equations are projected onto the space generated by the *centered* likelihood ratios.

Barankin's characterization of the elements of the space Λ_∞ as elements that are the *unique* LMVUE of their own expectations leads to the following completeness type result.

Proposition 2.2

For $g = g(\psi, \lambda_0, \mathbf{x}) \in \Lambda_\infty$ *and* $E_\lambda[g(\lambda_0)] = 0$ *for all* λ *and* λ_0 *together imply that* $g(\lambda_0) = 0$ *a.s.* $dP(\lambda_0)$*.*

Proof

Take any $g_1(\lambda_0) \in \Lambda_\infty$, and assume that $g \in \Lambda_\infty$ satisfies the condition $E_\lambda[g(\lambda_0)] = 0$ for all λ and λ_0. Then for any constant τ the function $g_1 + \tau g \in \Lambda_\infty$ is another unbiased estimator of $E_\lambda[g_1]$. By appropriate choice of τ see Rao (1973, p.317, 5a.2.1) this function can be made to have variance less than or equal to that of g_1, which contradicts the uniqueness of g_1 as the LMVUE of its own expectation. Consequently $g(\lambda_0) = 0$ a.s. $d\mathrm{P}(\lambda_0)$. ■

This property is reminiscent both of completeness and ancillarity. Completeness is obvious; ancillarity ideas come in through the fact that, in Lehmann's terminology, we have that the only first order ancillary function in Λ_∞ is the zero function.

The space Λ_∞ also has an important characterization when there exists a boundedly complete sufficient statistic for λ, say T. In this situation Λ_∞ is precisely the space of mean zero T-measurable functions. When the boundedly complete sufficient statistic depends on ψ, that is $T = T(\psi)$, then Λ_∞ is the space of mean zero $T(\psi)$-measurable functions.

The goal will be to select an element from $\mathcal{G}_0^r$ that has maximal *information* at λ_0, where the information in an unbiased estimating equation g is

$$I_g = E[\nabla_\psi g(\psi,\lambda)]\, E[g(\psi,\lambda)g(\psi,\lambda)^T]^{-1}\, E[\nabla_\psi g(\psi,\lambda)]^T.$$

Note that the information is a matrix-valued quantity, so that some care needs to be taken when it is said to be "maximized". It will be assumed that the matrix $E[g(\psi,\lambda)g(\psi,\lambda)^T]$ is of full rank, and hence invertible.

The information is invariant under the multiplication of an estimating function g, by a scalar $c(\psi,\lambda_0)$. It therefore makes sense to partition $\mathcal{G}_0$ according to the equivalence relation, $g_1 \sim g_2$ if and only if there exists a non-zero function $c(\psi,\lambda_0)$ such that $g_1 = c\, g_2$. Henceforth, the identification of any element of $\mathcal{G}_0$ will be taken to be an identification only up to the equivalence class that contains the function.

By construction, the matrix I_g is a non-negative definite matrix. It is sometimes possible to order a set of symmetric matrices according to the Löwner order, here denoted by $\stackrel{L}{\geq}$. A symmetric matrix R is said to be greater than or equal to the symmetric matrix S according to the Löwner order, and written $R \stackrel{L}{\geq} S$, if the matrix $R-S$ is nonnegative definite. R is said to be strictly greater than S if the difference $R-S$ is positive definite; this is written $R \stackrel{L}{>} S$. If $R \stackrel{L}{\geq} S$ then it can be shown that

$\lambda_i \geq \mu_i$ for $i = 1, 2 \cdots r$ where λ_i and μ_i are the eigenvalues of the matrices R and S respectively. If $R \overset{L}{>} S$ then strict inequality exists for at least one pair (λ_i, μ_i). Consequently if $R \overset{L}{\geq} S$ and $\phi(R)$ is any monotonically increasing function of the eigenvalues of R, then it immediately follows that $\phi(R) \geq \phi(S)$. Three commonly encountered monotonic functions of the eigenvalues used to measure information are

1. $\mathrm{tr}(R) = \sum \lambda_i$.

2. $\det(R) = \prod \lambda_i$.

3. The minimum eigenvalue of R, $m(R) = min\{\lambda_i\}$.

(See, Pukelsheim (1987) for a discussion of *information functionals* in the context of optimal design.)

In general, because the Löwner order is only a partial order, a set of symmetric matrices need not contain a maximum or minimum element. However, in this information matrix setting there is a *unique* maximal element, it is the information matrix for the score function. The maximality of this largest element will be preserved under any ordering that is a monotonic function of the original ordering. The important consequence is that maximality with respect to any one information ordering is equivalent to maximality with respect to any of the others. This result is discussed in Chandrasekar and Kale (1984) for the equivalence of Löwner, trace and determinant optimality.

Since λ_0 is for the time being assumed known, any optimality claims in the following lines are for local optimality, unless stated otherwise. To obtain the first result a simple result is needed that generalizes the familiar Cauchy-Schwarz inequality:

Proposition 2.3 *For any two estimating equations g_1 and $g_2 \in \mathcal{G}_0^r$*

$$E[g_1 g_1^T] - E[g_1 g_2^T]E[g_2 g_2^T]^{-1}E[g_2 g_1^T] \overset{L}{\geq} 0$$

with equality if and only if g_2 is a linear function of g_1.

This result be used to establish the optimality of the score function within the class $\mathcal{G}_0^r$.

Theorem 2.1 *Let U_0 denote the score function for the parameter of interest, that is $U_0 = \nabla_\psi l(\psi, \lambda_0)$. Assuming the basic regularity conditions hold, then the maximal element in $\mathcal{G}_0^r$ with respect to the Löwner order is U_0.*

Proof

For any estimating function $g \in \mathcal{G}_0^r$ we have $E[\nabla_\psi g] = -E[gU_0^T]$. It follows that the G-information, I_g, is

$$I_g = E[g(\psi, \lambda_0)U_0(\psi, \lambda_0)^T]\, E[g(\psi, \lambda)g(\psi, \lambda)^T]^{-1}\, E[g(\psi, \lambda)U_0(\psi, \lambda_0)]^T,$$

so that by Proposition 2.3

$$I_g \overset{L}{\leq} E[U_0(\psi, \lambda_0)U_0(\psi, \lambda_0)^T].$$

The identity $E[\nabla_\psi U_0] = -E[U_0 U_0^T]$, establishes that $I_{U_0} = E[U_0 U_0^T]$. Hence

$$I_g \overset{L}{\leq} I_{U_0} \text{ for all } g \in \mathcal{G}_0^r$$

as required. ∎

Note that this establishes that U_0 is a maximal element in $\mathcal{G}_0^r$ with respect to the Löwner order. Since U_0 is unique, and all other elements of $\mathcal{G}_0^r$ are comparable to it in the Löwner sense, the equivalence of the information optimality criteria is immediate.

Remark: If the problem in fact contained no nuisance parameters, so that λ_0 is vacuous, then the score function $U_0(\psi, \lambda_0) = U_0(\psi)$ is evidently globally optimal, which is Godambe's (1960) result.

The following theorem provides a basic result which essentially states, that if there exist two nested subspaces, say $\mathcal{H}^r \subseteq \mathcal{H}_0^r$, where $\mathcal{H}^r = \overset{r}{\times}\, \mathcal{H}$, and U_0 is optimal with respect to the Löwner order on the information matrices in the wider class, then the projection of U_0 onto $\mathcal{H}^r$ is optimal in the restricted class $\mathcal{H}^r$.

Theorem 2.2 *Let* $\overset{r}{\times}\, \mathcal{H} = \mathcal{H}^r \subseteq \mathcal{H}_0^r$*, and let projection onto the space* $\mathcal{H}^r$ *be denoted by* Π_1*. Then*

$$I_{\Pi_1 U_0} \overset{L}{\geq} I_g \text{ for all } g \in \mathcal{H}^r.$$

Proof

If $g \in \mathcal{H}_0^r$ is the vector valued function $g = (g_1, g_2 \cdots g_r)^T$, then clearly $\Pi_1 g = (\Pi g_1, \Pi g_2 \cdots \Pi g_r)^T$, where Π denotes projection of a function in $\mathcal{H}_0$ onto the subspace $\mathcal{H}$. This follows because both $\mathcal{H}^r$ and $\mathcal{H}_0^r$ are defined as Cartesian products.

The proof makes use of the well known fact about projections, that if $g_1 \in \mathcal{H}$ and $g_0 \in \mathcal{H}_0$ then

$$\langle g_1, g_0 \rangle = \langle g_1, \Pi g_0 \rangle. \tag{2.9}$$

Let $g \in \mathcal{H}^r$, then by definition

$$\begin{aligned} I_g &= E[\nabla_\psi g]\, E[gg^T]^{-1}\, E[\nabla_\psi g]^T \\ &= E[gU_0^T]\, E[gg^T]^{-1}\, E[gU_0^T]^T. \end{aligned}$$

Each element of the matrix $E[gU_0^T]$ is an inner product so that it can by (2.9) be written as $E[g(\Pi_1 U_0)^T]$. Hence,

$$I_g = E[g(\Pi_1 U_0)^T]\, E[gg^T]^{-1}\, E[g(\Pi_1 U_0)^T]^T. \tag{2.10}$$

But by applying Proposition 2.3 to (2.10) it follows that

$$I_g \overset{L}{\leq} E[\Pi_1 U_0 (\Pi_1 U_0)^T] = E[\nabla_\psi \Pi_1 U_0]\, E[\Pi_1 U_0 (\Pi_1 U_0)^T]^{-1}\, E[\nabla_\psi \Pi_1 U_0]^T = I_{\Pi_1 U_0}$$

as required. ■

The particular subspace of interest of $\mathcal{G}_0^r$ is the subspace of functions satisfying the strong unbiasedness condition

$$E_{\psi,\lambda}[g(\psi, \lambda_0)] = 0 \text{ for all } \theta, \lambda_0. \tag{2.11}$$

This subspace will be denoted by $\mathcal{G}_\infty^r$. Under assumptions equivalent to those of Barankin (1949) on the likelihood ratios, $\mathcal{G}_\infty^r$ is in fact orthogonal to Λ_∞^r, and the following theorem states that the direct sum of these subspaces is equal to $\mathcal{G}_0^r$.

Theorem 2.3 *The orthogonal complement of $\mathcal{G}_\infty^r \subseteq \mathcal{G}_0^r$ is Λ_∞^r, so that $\mathcal{G}_\infty^r \oplus \Lambda_\infty^r = \mathcal{G}_0^r$.*

Proof The fact that $\mathcal{G}_\infty \oplus \Lambda_\infty = \mathcal{G}_0$ will be proved, from which the main result will immediately follow. It is enough to show that $\mathcal{G}_\infty = \Lambda_\infty^\perp$, since then, see (Halmos, 1951, §12)

$$\mathcal{G}_0 = \mathcal{G}_\infty \oplus \mathcal{G}_\infty{}^\perp = \mathcal{G}_\infty \oplus \left(\Lambda_\infty^\perp\right)^\perp = \mathcal{G}_\infty \oplus \Lambda_\infty.$$

Let $g \in \Lambda_\infty^\perp$, that is $\langle g, h\rangle = 0$ for all $h \in \Lambda_\infty$ This is equivalent to g being orthogonal to all the generators of Λ_∞, so that

$$E_{\lambda_0}\left[g\left(\frac{p(\lambda)}{p(\lambda_0)} - 1\right)\right] = 0 \text{ for all } \lambda$$

that is $E_\lambda[g] - E_{\lambda_0}[g] = 0$ for all λ. But $g \in \mathcal{G}_0$ so that $E_{\lambda_0}[g] = 0$ and hence $E_\lambda[g] = 0$ for all λ so that $g \in \mathcal{G}_\infty$, which establishes $\Lambda_\infty^\perp \subseteq \mathcal{G}_\infty$.

Conversely assume that $g \in \mathcal{G}_\infty$. Then $E_\lambda[g(\lambda_0)] = 0$ for all λ and λ_0. Clearly $[g(\lambda_0) - \Pi g(\lambda_0) + \Pi g(\lambda_0)] \in \mathcal{G}_\infty$, but $\Pi g(\lambda_0) \in \Lambda_\infty$ by the definition of projection, so that $g(\lambda_0) - \Pi g(\lambda_0) \in \Lambda_\infty^\perp$. Applying the fact that $\Lambda_\infty^\perp \subseteq \mathcal{G}_\infty$ it follows that $E_\lambda[g(\lambda_0) - \Pi g(\lambda_0)] = 0$ and hence that $E[\Pi g(\lambda_0)] = 0$ for all λ and λ_0. Since $\Pi g(\lambda_0) \in \Lambda_\infty$ and $E[\Pi g(\lambda_0)] = 0$ for all λ Proposition 2.2 applies, so that $\Pi g(\lambda_0) = 0$. That is, Π projects any function in $\mathcal{G}_\infty$ onto the zero function. Consequently $g \in \Lambda_\infty^\perp$ and hence $\mathcal{G}_\infty \subseteq \Lambda_\infty^\perp$. This together with $\Lambda_\infty^\perp \subseteq \mathcal{G}_\infty$ establishes the result. ■

The next theorem is an immediate consequence of Theorem 2.3.

Theorem 2.4 *The optimal estimating equation within the class $\mathcal{G}_\infty^r$ is*

$$U_\infty = U_0 - \Pi_\infty U_0$$

where Π_∞ denotes projection onto the subspace Λ_∞^r.

Proof

U_0 can be written as

$$U_0 = U_0 - \Pi_\infty U_0 + \Pi_\infty U_0 \tag{2.12}$$

Clearly $\Pi_\infty U_0 \in \Lambda_\infty^r$, and $U_0 - \Pi_\infty U_0$ is orthogonal to $\Pi_\infty U_0$, and hence by Theorem 2.3 an element of $\mathcal{G}_\infty^r$. Thus (2.12) represents the decomposition of U_0 into a part in $\mathcal{G}_\infty^r$ and a part in Λ_∞^r, so that by the definition of projection $U_0 - \Pi_\infty U_0$ is the projection of U_0 onto $\mathcal{G}_\infty^r$ and consequently by Theorem 2.2 the optimal estimating function within $\mathcal{G}_\infty^r$. ■

Corollary 2.1 *Let $\mathcal{G}_\infty^r$ be the subspace of all estimating functions g satisfying (2.11) and denote projection onto $\mathcal{G}_\infty^r$ by Π_1. If there exists a boundedly complete sufficient statistic for λ, say $T(\psi)$ for each λ, then the optimal estimating equation in $\mathcal{G}_\infty^r$ is $U_c(\psi, \lambda_0)$.*

Proof

Write U_0 as

$$U_0 = U_0(\psi, \lambda_0) - E_\psi[U_0(\psi, \lambda_0) \,|\, T(\psi)] + E_\psi[U_0(\psi, \lambda_0) \,|\, T(\psi)].$$

Clearly $U_c = U_0(\psi, \lambda_0) - E_\psi[U_0(\psi, \lambda_0) \,|\, T(\psi)] \in \mathcal{G}^r_\infty$, and $E_\psi[U_0(\psi, \lambda_0) \,|\, T(\psi)]$ is a $T(\psi)$ measurable function, and hence a member of Λ^r_∞. Therefore U_c is precisely the projection of U_0 onto $\mathcal{G}^r_\infty$ and so by Theorem 2.2 must be the optimal estimating equation. ∎

Corollary 2.2 *If there exists a boundedly complete sufficient statistic for λ, say T, that is a function of the data alone, so that the density factors as*

$$p(\mathbf{x}, \theta) = f_{\mathbf{X}|T}(\mathbf{x}; t, \psi)\, h(t, \psi, \lambda)$$

then the optimal estimating equation within the class $\mathcal{G}^r_\infty$ is $U_c = \frac{\partial}{\partial\psi} ln\, f_{\mathbf{X}|T}(\mathbf{x}; t, \psi)$.

Proof

By assumption the score U_0 factorizes as

$$U_0 = \frac{\partial}{\partial\psi} ln\, f_{\mathbf{X}|T}(\mathbf{x}; t, \psi) + \frac{\partial}{\partial\psi} ln\, h(t, \psi, \lambda_0). \tag{2.13}$$

Since the first part is totally independent of λ it is clearly an element of $\mathcal{G}^r_\infty$. The second part is T measurable, and hence an element of Λ^r_∞. ∎

Remark: The function $\frac{\partial}{\partial\psi} ln\, f_{\mathbf{X}|T}(\mathbf{x}; t, \psi)$ does not depend on λ_0 and is therefore a uniformly optimal estimating equation. This is Godambe's (1976) result.

Remark: It may at first seem surprising that the conditional score in the Type II model depends on λ_0, but if T is a function of ψ, then the derivative of $T(\psi)$ need not be $T(\psi)$ measurable, so that the function $\frac{\partial}{\partial\psi} ln\, h(t, \psi, \lambda_0)$ in (2.13) will not lie in Λ^r_∞ unless T is free of ψ.

2.4 Approximations to U_∞ and the Bhattacharyya basis

Recall that the optimal estimating equation within the class $\mathcal{G}_\infty^r$ is given by U_0 with its projection onto the space Λ_∞^r removed. It may be difficult from a technical (if not conceptual) point of view to carry out this projection. In a similar vein, even in a conditional problem where the optimality of the conditional score is assured, it might not be feasible to construct it due to the numerical intractability of the relevant conditional distributions. This leads to the idea of approximating U_c or U_∞.

One natural way of doing this is to construct a basis for the space Λ_∞^r, and rather than project onto the space Λ_∞^r, project onto a successive sequence of subspaces of Λ_∞^r.

Alternatively, one might consider the strong unbiasedness condition that defines the class $\mathcal{G}_\infty^r$, too extreme. In this case it makes sense to relax the unbiasedness condition in a progressive manner to allow consideration of a wider class of estimating functions.

These two approaches will be reconciled in the development that follows. More notation and definitions need to be introduced.

It will now be assumed that the data observed $\mathbf{x}$, is a realization of n independent random variables, with densities, $p_1(\theta), p_2(\theta) \cdots p_n(\theta)$ respectively. It follows that the density of $\mathbf{X}$ is by independence the product of the individual densities. As a matter of notational convenience, the elements of a vector will be indexed by superscripts, so that for example if λ is a vector, then λ^2 denotes the second element of the vector, and not λ squared.

Define the function $V_{i_1,i_2\cdots i_t}$ as

$$V_{i_1 i_2\cdots i_t}=\frac{\frac{\partial^t}{\partial\lambda^{i_1}\partial\lambda^{i_2},\cdots,\partial\lambda^{i_t}}\prod_{i=1}^n p_i(\mathbf{x},\psi,\lambda)}{\prod p_i(\mathbf{x},\psi,\lambda)},$$

where $1\le i_j\le n, j=1,2\cdots s$ (recall that s denotes the number of nuisance parameters). For example if there were 3 observations and two nuisance parameters, so that $\lambda=(\lambda^1,\lambda^2)$ then the function V_{12} would be

$$V_{12}=\frac{\frac{\partial^2}{\partial\lambda^1\partial\lambda^2}\prod p_i(\mathbf{x},\psi,\lambda)}{\prod p_i(\mathbf{x},\psi,\lambda)}. \tag{2.14}$$

In this notation the score vector for the nuisance parameters are denoted simply by V_i.

The $V_{i_1 i_2\cdots i_t}$ will be referred to as the *Bhattacharyya Basis* on account of their use by A. Bhattacharyya (1946) in his investigation of unbiased estimation and the sharpening of the Cramér-Rao lower bound.

The following regularity conditions associated with the densities will be assumed

- $V_{i_1 i_2\cdots i_k}$ exists and is finite a.s. P_θ for all θ and for $k=1,2,\cdots,t$.
- $V_{i_1 i_2\cdots i_k}\in\mathcal{L}_2$ for all θ and for $k=1,2,\cdots,t$.
- For any function $g\in\Lambda_\infty$,

$$\begin{aligned}\frac{\partial^t}{\partial\lambda^{i_1}\partial\lambda^{i_2},\cdots,\partial\lambda^{i_t}}E_\lambda[g(\lambda_0)] &= \frac{\partial^t}{\partial\lambda^{i_1}\partial\lambda^{i_2},\cdots,\partial\lambda^{i_t}}\int_\Omega g(\lambda_0)p(\lambda)dx\Big|_{\lambda=\lambda_0}\\ &= \int_\Omega g(\lambda_0)\,\frac{\partial^t}{\partial\lambda^{i_1}\partial\lambda^{i_2},\cdots,\partial\lambda^{i_t}}p(\lambda)\Big|_{\lambda=\lambda_0}dx\\ &= E_{\lambda_0}[g(\lambda_0)V_{i_1 i_2\cdots i_t}].\end{aligned}$$

These regularity conditions will be denoted by R(t), and if they hold for all t, then the conditions will be written as R(∞) .

The class of functions $\{V_{i_1,i_2\cdots i_t} : 1 \le i_1 \le s, 1 \le i_2 \le s, \cdots 1 \le i_t \le s\}$ will be denoted by $\mathbf{V}_t$ and can be differentiated from the score vector V_i by the use of the bold V, $\mathbf{V}$.

Let Λ_1 denote the closed linear manifold generated by $\{\mathbf{V}_1\}$, Λ_2 denote the closed linear manifold generated by $\{\mathbf{V}_1, \mathbf{V}_2\}$ and in general Λ_t is the subspace generated by $\{\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_t\}$. Clearly $\{\Lambda_i : i = 1, 2, \cdots\}$ is an increasing sequence of subspaces, and the claim is that under appropriate regularity conditions, the limit of this sequence, that is $\bigvee_t \mathbf{V}_t$ will equal Λ_∞, the space generated by the centered likelihood ratios.

The proof is similar to that of Mase (1977). The idea is to show that the only element of Λ_∞ that is orthogonal to all of the $V_{i_1 i_2 \cdots i_t}$ is the zero function, so that the $\{V_{i_1,i_2\cdots i_t}\}$ generate Λ_∞.

Theorem 2.5 *Assume the regularity conditions $R(\infty)$ hold. Furthermore, assume that for any function $g(\lambda_0) \in \Lambda_\infty$ the function $E_\lambda[g(\lambda_0)]$ is analytic at λ_0. Then $\bigvee_t \mathbf{V}_t = \Lambda_\infty$.*

<u>Proof</u>

Assume that there exists an element, say g, of Λ_∞ that is orthogonal to all the $V_{i_1 i_2 \cdots i_t}$. That is

$$E_{\lambda_0}[g(\lambda_0) V_{i_1 i_2 \cdots i_t}] = 0 \text{ for all } 1 \le i_j \le s \text{ and } t = 1, 2 \cdots$$

Since $E_\lambda[g(\lambda_0)]$ is assumed analytic at λ_0, it has a convergent Taylor series expansion about λ_0,

$$E_\lambda[g(\lambda_0)] =$$
$$E_{\lambda_0}[g(\lambda_0)] + \sum_{k=1}^{\infty} \frac{1}{k!} \sum_{i_1=1}^{s}, \cdots, \sum_{i_k=1}^{s} \frac{\partial^k}{\partial\lambda^{i_1}\partial\lambda^{i_2}, \cdots, \partial\lambda^{i_k}} E_\lambda[g(\lambda_0)]\Big|_{\lambda=\lambda_0} \cdot \prod_{j=1}^{k} (\lambda^{i_j} - \lambda_0^{i_j}).$$

See Serfling (1980, p.44, Theorem B) for a multivariate Taylor series formula. The first term in the above series is 0 since $g \in \mathcal{G}_0$. By the assumed regularity R(∞) the terms $\frac{\partial^k}{\partial\lambda^{i_1}\partial\lambda^{i_2},\cdots,\partial\lambda^{i_k}} E_\lambda[g(\lambda_0)]\Big|_{\lambda=\lambda_0}$ are equal to $E_{\lambda_0}[g(\lambda_0)V_{i_1 i_2 \cdots i_t}]$, but by the orthogonality assumption this is zero. Hence $E_\lambda[g(\lambda_0)]$ is identically equal to 0. Consequently $g \in \mathcal{G}_\infty$. But $g \in \mathcal{G}_\infty$ and $g \in \Lambda_\infty$ by Proposition 2.2, together imply that $g(\lambda_0)$ is identically equal to zero dP(λ_0), which gives the result. ■

The regularity assumptions appear severe. But as Blight and Rao (1974), Mase (1977) and Ghosh and Sathe (1987) all point out, the conditions are met in canonically parameterized exponential families. See Mase (1977) for a wider listing of distributions satisfying the assumptions.

2.5 The family of projective scores, their construction and information properties

In this section a series of levels of unbiasedness for estimating equations will be introduced that fill the gap between the weak form of unbiasedness satisfied by the elements of $\mathcal{G}_0$ and the strong form satisfied by the elements of $\mathcal{G}_\infty$. An estimating function $g(\psi, \lambda_0)$ that meets the regularity conditions R(t) will be termed $\mathcal{G}_t$*-unbiased* if it satisfies:

$$\frac{\partial^k}{\partial\lambda^{i_1}\partial\lambda^{i_2},\cdots,\partial\lambda^{i_k}} E_\lambda[g(\lambda_0)]\Big|_{\lambda=\lambda_0} = 0 \text{ for all } \psi, \lambda \text{ and } \lambda_0, k = 1, 2\cdots t. \tag{2.15}$$

Note that by the assumed regularity

$$\frac{\partial^k}{\partial\lambda^{i_1}\partial\lambda^{i_2},\cdots,\partial\lambda^{i_k}} E_\lambda[g(\lambda_0)]\Big|_{\lambda=\lambda_0} = E_{\lambda_0}[g(\lambda_0)\; V_{i_1 i_2 \cdots i_t}|_{\lambda=\lambda_0}]$$

so that an equivalent way of stating that a function is $\mathcal{G}_t$-unbiased or an element of $\mathcal{G}_t$ is to assert that it is orthogonal to the space Λ_t. It can be shown in a manner similar to that of Theorem 2.3 that $\mathcal{G}_t \oplus \Lambda_t = \mathcal{G}_0$. Writing the Cartesian product spaces as $\mathcal{G}_t^r$ and Λ_t^r it follows from Theorem 2.2 that

Theorem 2.6 *The optimal estimating equation in the class of $\mathcal{G}_t^r$-unbiased functions is*

$$U_t = U_0 - \Pi_t U_0$$

where Π_t denotes projection onto the space Λ_t^r.

U_t will be called the t^{th}*-projective score*. Furthermore, if the conditions of Theorem 2.5 hold then $\lim_{t\to\infty} \Pi_t U_0 = \Pi_\infty U_0$. This follows because the limit of the projection Π_t is the projection onto the space $\bigvee \mathbf{V}_t = \Lambda_\infty$.

Since the t^{th}-projective score can be constructed by taking the score function U_0 and removing its projection onto the space Λ_t^r, in other words U_0 is made orthogonal to Λ_t^r, it can be constructed by regressing out the functions $V_{i_1,i_2,\cdots,i_t}$ from each component of the vector U_0. The projection of U_0 onto Λ_t^r can be characterized as the point in Λ_t^r that satisfies the relation

$$\langle U_0 - P_t U_0, P_t U_0 \rangle = 0.$$

If the entire set of functions $V_{i_1,i_2,\cdots,i_k}$, $k = 1, 2\cdots t$, listed as a vector is denoted by $\mathbf{V}$ then a matrix of coefficients A is needed that satisfy the relation

$$\mathrm{tr}E[(U_0 - A\mathbf{V})\mathbf{V}^T A^T] = 0,$$

and it is easily verified that $A = E[U_0\mathbf{V}^T]E[\mathbf{V}\mathbf{V}^T]^+$ is a solution. The symbol $^+$ denotes the Moore-Penrose generalized inverse. Consequently

$$U_t = U_0 - E[U_0\mathbf{V}^T]E[\mathbf{V}\mathbf{V}^T]^+\mathbf{V}. \tag{2.16}$$

It would at first appear expedient if the matrix $E[\mathbf{V}\mathbf{V}^T]$ were assumed non-singular, so that its inverse existed and was unique. This could of course be achieved by removing redundant elements from the set $\mathbf{V}$. There are potentially many of these redundant elements; for example in a model with two nuisance parameters λ^1 and λ^2, the functions V_{12} and V_{21} are identical, since the order of differentiation can be exchanged. But it will prove to be very useful to allow the repetition of elements in the set $\mathbf{V}$, and hence for $E[\mathbf{V}\mathbf{V}^T]$ to be non-singular. This is because the induced high order symmetries on $E[\mathbf{V}\mathbf{V}^T]$ can be exploited and great simplifications in practical computations are achieved. In Chapters 4 and 7, the structural properties of a non-singular $E[\mathbf{V}\mathbf{V}^T]$ matrix will be fully exploited.

As an aside, it is useful to consider the need for the Moore-Penrose inverse in (2.16). The Moore-Penrose inverse G of a matrix A satisfies four conditions,

$$\begin{aligned} AGA &= A \\ GAG &= G \\ (AG)^* &= AG \\ (GA)^* &= GA \end{aligned}$$

which in fact ensure its uniqueness (the * denotes complex conjugate). For the proofs presented in this thesis it is only necessary for the inverse to satisfy the first two conditions (providing an inverse that is sometimes called a g_2-inverse (Pringle & Rayner, 1971)), but the Moore-Penrose inverse is preferred on account of its uniqueness and the multitude of theorems available for its analysis.

The information in U_t, I_{U_t}, has a simple form:

Theorem 2.7

$$I_{U_t} = E[U_0 U_0^T] - E[U_0 \mathbf{V}^T] E[\mathbf{V}\mathbf{V}^T]^+ E[\mathbf{V} U_0^T].$$

<u>**Proof**</u>

By definition

$$I_{U_t} = E[\nabla U_t]E[U_t U_t^T]^{-1}E[\nabla U_t]^T. \tag{2.17}$$

But from the assumed regularity on U_t, we have that

$$\begin{aligned} E[\nabla U_t] &= -E[U_t U_0^T] \\ &= -E[U_t(U_0 - \Pi_t U_0 + \Pi_t U_0)^T] \\ &= -E[U_t U_t^T] - E[U_t(\Pi_t U_0)^T] \\ &= -E[U_t U_t^T]. \end{aligned} \tag{2.18}$$

as U_t is by definition orthogonal to Λ_t^r and hence to $\Pi_t U_0$. By substituting (2.18) into (2.16) it follows that $I_{U_t} = E[U_t U_t^T]$, but

$$E[U_t U_t^T] = E[(U_0 - E[U_0\mathbf{V}^T]E[\mathbf{V}\mathbf{V}^T]^+\mathbf{V})(U_0 - E[U_0\mathbf{V}^T]E[\mathbf{V}\mathbf{V}^T]^+\mathbf{V})^T],$$

which on expansion gives the result. ■

In some special problems the approximations, U_t to the conditional score function, U_c, will be exact for some finite value of t. Lindsay and Waterman (1991, p. 23) give a series of basic examples that illustrate this possibility.

As an illustration of the ability of the projective scores to approximate the conditional score, in terms of their information content, we introduce the Binomial extension to Example 1:

Example 1 *Paired Binomial*

The experiment consists of the allocation of two treatments to two groups of r_i and s_i individuals respectively, so that one group obtains treatment 1 say, and the other group receives treatment 2. A binary response is observed on each individual, and the overall response for a group is the sum of the binary responses. A series of n independent pairs of groups is observed, and the pairs are matched according to some characteristic. For the i^{th} group let X_{i1} be the sum of the binary responses of the individuals that receive treatment 1, and let X_{i2} be the sum of the binary responses of the individuals that receive treatment 2. A common model for this situation is to hypothesize

$$X_{i1} \sim Bin\left(r_i, \frac{\exp(\psi + \lambda_i)}{1 + \exp(\psi + \lambda_i)}\right) \qquad X_{i2} \sim Bin\left(s_i, \frac{\exp(\lambda_i)}{1 + \exp(\lambda_i)}\right).$$

The parameter of interest is the common odds ratio ψ, the nuisance parameters being the λ_i. If $r_i = s_i = 1$ for all i then this is exactly the model described in Example 1.

Information comparisons between the projective and conditional scores were made for a single stratum, so that the subscript i is superfluous. A complete sufficient statistic for the nuisance parameter λ is $X_1 + X_2$. Since this is not a function of ψ this is a Type I problem, and a globally optimal estimating equation exists for ψ, being the conditional score U_c, where

$$U_c = X - \frac{\sum_{i=max(0,t-s)}^{min(r,t)} \binom{r}{i}\binom{s}{t-i} i\, exp(\psi\, i)}{\sum_{i=max(0,t-s)}^{min(r,t)} \binom{r}{i}\binom{s}{t-i} exp(\psi\, i)},$$

and $t = x_1 + x_2$. (See Cox and Snell (1989) for a full discussion of this model.)

The score functions for ψ and λ are

$$\begin{aligned} U_0 &= -r\frac{exp(\psi + \lambda)}{1 + exp(\psi + \lambda)} + x_1. \\ V_1 &= -r\frac{exp(\psi + \lambda)}{1 + exp(\psi + \lambda)} + x_1 - s\frac{exp(\lambda)}{1 + exp(\lambda)} + x_2. \end{aligned}$$

The calculation of V_{11} is facilitated by the use of a recurrence relation that will be described in Theorem 3.1.

The U_1 and U_2 informations were calculated, and compared to the information in the conditional score function $E[U_c^2] = -E[\frac{\partial}{\partial\psi}U_c]$. Table 2.2 shows the various informations for $\psi = 0.5$ and $\lambda = -0.25$. These parameter values correspond to probabilities of success in the two binomials of 0.562 and 0.438 respectively. The ratio of the square root of the U_1 and U_2 information to the conditional score information is also presented, this transformation being appropriate for the scale of the parameter estimates. The sample sizes, r and s have been set equal, consequently the experiment is equivalent to observing r independent Bernoulli pairs (X_{1j}, X_{2j}) so that the U_1 information increases linearly in n.

Two cases of note are when ψ equals 0 so that U_1 equals U_c (this follows from a completeness argument), and the case $r = s = 1$ where the $U_2 = U_c$ and so the I_{U_2} information is equal to the conditional information. To understand why, note that

$$U_2 = U_0 - \Pi_2 U_0.$$

where Π_2 is the projection onto the space spanned by $\{V_1, V_{11}\}$, that is Λ_2. V_1 is a linear function in t, and using Theorem 3.1, which states that $V_{11} = V_1' + V_1^2$, it follows that V_{11} is a quadratic in t. Hence Λ_2 is also the space spanned by mean zero quadratic functions of t. Now, U_c equals

$$\begin{aligned} U_0 - \Pi_\infty U_0 &= U_0 - (E[X_1|t] - E[X_1]) && (2.19)\\ &= U_0 - \left(\frac{\sum_{i=max(0,t-1)}^{min(1,t)} \binom{1}{i}\binom{1}{t-i} i\, exp(\psi\, i)}{\sum_{i=max(0,t-1)}^{min(1,t)} \binom{1}{i}\binom{1}{t-i} exp(\psi\, i)} - E[X] \right). && (2.20) \end{aligned}$$

Noting that t can only take on the values 0, 1 and 2 it follows that

$$\left(\frac{\sum_{i=max(0,t-1)}^{min(1,t)} \binom{1}{i}\binom{1}{t-i} i\, exp(\psi\, i)}{\sum_{i=max(0,t-1)}^{min(1,t)} \binom{1}{i}\binom{1}{t-i} exp(\psi\, i)} - E[X_1] \right)$$

can only take on three possible values and can hence be expressed as a quadratic in t. Therefore $\Pi_\infty U_0 = \Pi_2 U_0$, so that $U_c = U_2$.

Table 2.2: Comparison of projective and conditional informations in a paired Binomial experiment.

$r=s$	$U_1 Info$	$U_2 Info$	$U_c Info$	$\sqrt{\frac{U_1 Info}{U_c Info}}$	$\sqrt{\frac{U_2 Info}{U_c Info}}$
1	0.123067041	0.119318866	0.119318866	1.015585102	1.000000000
2	0.241340823	0.243583512	0.243581951	1.005225102	1.000003203
3	0.369201124	0.366896056	0.366895706	1.003136869	1.000000476
4	0.492268165	0.490068939	0.490068784	1.002241439	1.000000158
5	0.615335207	0.613194945	0.613194858	1.001743723	1.000000071
6	0.738402248	0.736299571	0.736299515	1.001426888	1.000000038
7	0.861469290	0.859392660	0.859392621	1.001207490	1.000000023
8	0.984536331	0.982478817	0.982478789	1.001046570	1.000000015
9	1.107603372	1.105560485	1.105560463	1.000923498	1.000000010
10	1.230670414	1.228639079	1.228639061	1.000826326	1.000000007
11	1.353737455	1.351715475	1.351715461	1.000747657	1.000000005
12	1.476804496	1.474790247	1.474790235	1.000682664	1.000000004
13	1.599871538	1.597863784	1.597863774	1.000628068	1.000000003
14	1.722938579	1.720936359	1.720936351	1.000581557	1.000000002
15	1.846005621	1.844008172	1.844008165	1.000541461	1.000000002
16	1.969072662	1.967079370	1.967079363	1.000506536	1.000000002
17	2.092139703	2.090150064	2.090150059	1.000475844	1.000000001
18	2.215206745	2.213220341	2.213220336	1.000448659	1.000000001
19	2.338273786	2.336290269	2.336290265	1.000424412	1.000000001
20	2.461340827	2.459359901	2.459359897	1.000402652	1.000000001

If the sample size is increased to say $r = s = n$, then $\Pi_\infty U_0$ is a function of t, which takes on the values $t = 0, 1, \ldots, 2n$, and hence will be expressible as a polynomial function of degree $2n$ of t. Similarly, the space Λ_{2n} is equivalent to the space spanned by polynomial functions in t of degree $2\,n$, so that $U_c = U_{2n}$.

It is immediately apparent from the Table 2.2 that both the U_1 and U_2 functions have informations extremely close to that of the conditional score. The U_2 information approximates the conditional information so well for all sample sizes that any further approximation, for example U_3, would be unnecessary. (Note that the sequence of approximations is monotonic in information.)

In fact the approximations are so good that one might consider that the calculations have unwittingly been carried out for a special, or near special case. To examine this possibility some assymetric sample sizes were considered. In Table 2.3 the sample sizes have been set to 3 and 7, and the information is calculated for $p_1 = 0.1, p_2 = 0.9$ and $p_1 = 0.4, p_2 = 0.7$.

Table 2.3: Comparison of informations for assymetric sample sizes in paired Binomial experiment.

p_1, p_2	$U_1 Info$	$U_2 Info$	$U_c Info$	$\sqrt{\frac{U_1 Info}{U_c Info}}$	$\sqrt{\frac{U_2 Info}{U_c Info}}$
0.1,0.9	0.189000000	0.141907675	0.128272993	1.213845152	1.051805236
0.4,0.7	0.483287671	0.473736114	0.473566326	1.010211833	1.000179249

Table 2.3 demonstrates that the U_1 and U_2 informations are very good

approximations to the conditional information, and even towards the edge of the parameter space, $p_1 = 0.1$, $p_2 = 0.9$, the U_2 information is still a reasonable substitute for the conditional information in this very unbalanced situation, where U_1 fails to work so well.

To understand how a change in sample size effects the results, Figure 2.1 plots the standardized U_1 information and U_c information for various sample sizes ($r = s$), where standardization of information has been carried out with respect to the sample size. The parameter values used were once again $\psi = 0.5$ and $\lambda = -0.25$. The U_2 information is also plotted, but is so close to the U_c information that it is indistinguishable from it at the resolution of the graph!

From a practical standpoint, for these parameter values, by the time the sample size is about 5 the U_1 information is equivalent to the U_c information.

Figure 2.2 presents a graphical comparison of the U_1 and U_2 informations relative to the U_c information. As in the case of Table 2.3, the comparison is made in terms of the square root of the ratio of the informations. Sample sizes of $r = 3$ and $s = 7$ were used; p_1 and p_2 both ranged from 0.1 to 0.9. When the probabilities are equal the odds ratio ψ is 0, and the three informations are identical, hence the two surfaces equal 1 along the line $p_1 = p_2$. Notice that the worst relative informations occur for extreme asymmetry of p_1 and p_2, but even so the relative U_2 information does an extremely good job over the entire practical region of the parameter space. (Practical is used in the sense that there is little hope of drawing asymptotic inferences in binomial experiments, that have small sample sizes and extreme probabilities of success.)

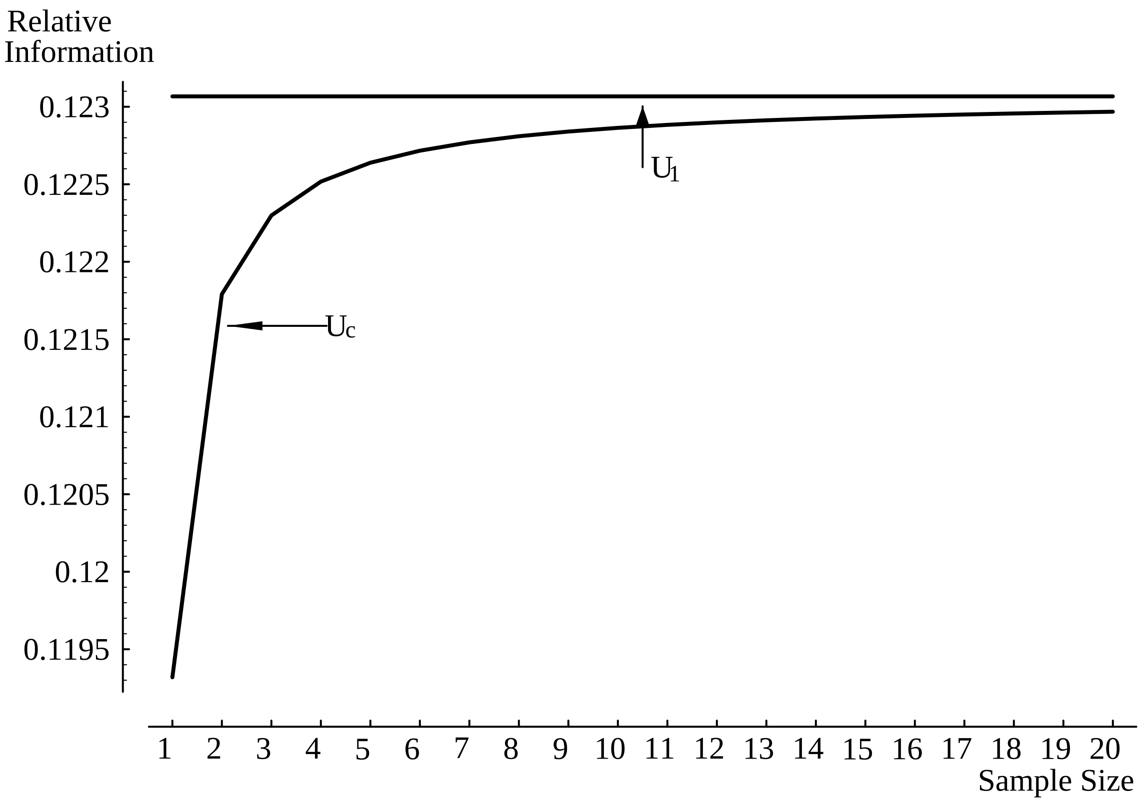

Figure 2.1: Standardized information comparison for a paired Binomial experiment.

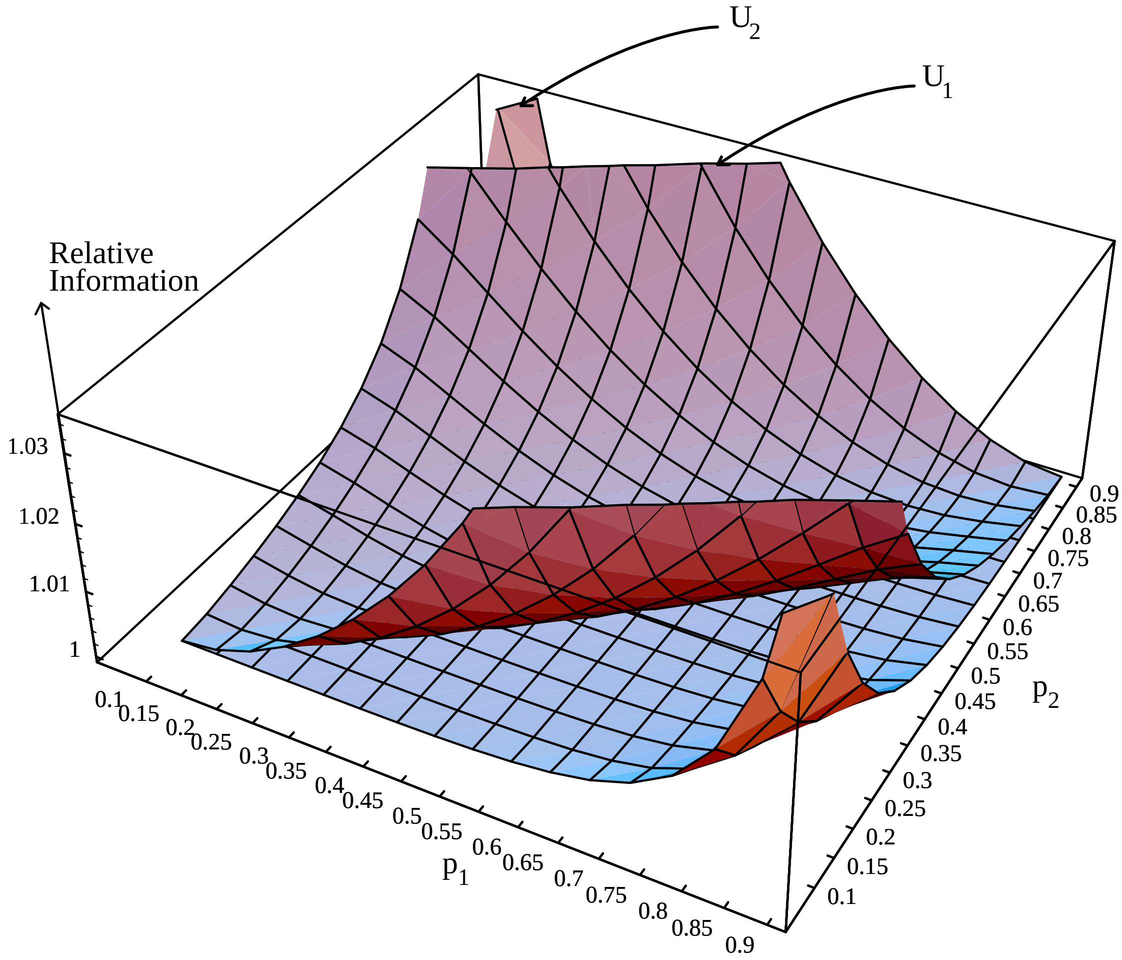

Figure 2.2: Relative information comparison for a paired Binomial experiment.