The accuracy of projected score methods in approximating conditional scores

By RICHARD P. WATERMAN
Department of Statistics, University of Pennsylvania,
Philadelphia, PA 19104, U.S.A.
and
BRUCE G. LINDSAY
Department of Statistics, Pennsylvania State University,
University Park, PA 16802, U.S.A.

Summary

This paper presents results that extend the projected score methods of Small & McLeish (1989). The main result shows that the conditional score function may be approximated with arbitrarily small stochastic error. The approximation method is pinned on a natural basis for the space of centered likelihood ratios. The desirability of using this basis is established. Constructive properties of the basis elements are presented that make higher order asymptotics tractable. In particular, a U-statistic representation theorem and expectation identities of a type similar to the Bartlett identities are presented. The results are applied to a canonically parameterized exponential family model where it is shown that the projected scores provide a particularly accurate approximation to the true conditional score function.

Some key words: Bhattacharyya basis; Conditional score; Exponential family; Information optimality; Projected score.

1 Introduction

The main contribution of this paper is to show that a particular statistical methodology, which we call the projected score approach, can be used to approximate conditional score functions with an arbitrarily small stochastic error. One potential value is clear: projected score methods thereby provide an algebraically straightforward approximation method for potential application in cases when the conditional score is not computationally tractable. However, unlike exact conditional methods, the projected score methodology is applicable in any smooth model, so these results also provide partial support for a larger claim: that projected score methodology is a natural extension of the idea of conditioning to models where exact conditional procedures fail to exist. In the process of establishing the approximation results, we will develop a toolbox of mathematical properties of the projected scores that will be useful for further research into their inferential properties.

The setting is a parametric statistical model $f(x; \psi, \lambda)$ which depends on ψ, the parameters of interest, and λ, a set of nuisance parameters. These will be taken to be r and s dimensional respectively. There are a variety of available strategies for dealing with nuisance parameter problems. The most straightforward case occurs when a complete sufficient statistic, say $S(X)$, exists for the nuisance parameter. This situation will be called a Type I problem. Following Fisher (1935) the standard analysis is conditional upon the value of this sufficient statistic. Among the theoretical justifications for conditioning is the optimality of conditional approaches in the theory of unbiased tests (Lehmann, 1986, Ch. 4). More germane here is the work of Godambe (1976) in estimating functions, who established that the conditional score function

$$U_c = U - E(U \mid S) \tag{1}$$

where $U = \partial l(\psi, \lambda; x)/\partial\psi$ is the ψ score and $l(\psi, \lambda; x)$ the loglikelihood, is optimal within a class of unbiased estimating equations. Note that in the Type I problem the conditional score U_c is just the derivative of the log conditional likelihood and so is no longer a function of λ. In a second type of conditional problem, which we call Type II, the complete and sufficient statistic depends on the true value of ψ, say $S(\psi; X)$, in which case the conditional score (1) typically depends on λ but still has some optimality properties (Lindsay, 1982). We will say a problem is Type III if neither of the above conditional structures exist.

The projected score methodology can be described in outline as follows: Let P_0 be the probability distribution determined by the fixed parameter values (ψ_0, λ_0). Let $\Lambda_{CLR}(P_0)$ be the closed linear subspace of $\mathcal{L}_2(P_0)$ generated by the set of centered likelihood ratio functions $[\{f(x;\psi_0,\lambda)/f(x;\psi_0,\lambda_0)\} - 1]$. One statistical interpretation of Λ_{CLR} is that it is the space of locally minimum variance unbiased estimators of functions of λ, for ψ_0 fixed (Barankin, 1949), centered to have mean zero under P_0. Next, if we let Π_{CLR} denote the $\mathcal{L}_2(P_0)$ projection operator onto Λ_{CLR} then the projected score function $U_{CLR} = U - \Pi_{CLR}U$ has the strong debiassing property $E_{\psi,\lambda}\{U_{CLR}(\psi,\lambda_0)\} = 0$ for all λ. In the terminology of Small & McLeish (1989), Λ_{CLR} is an example of a complete E-sufficient subspace.

Although a general statistical interpretation of U_{CLR} is not obvious, it has a clear meaning in a Type I or II problem. In these cases, the minimum variance unbiased estimators are all functions of S, so that Λ_{CLR} equals the space of all finite variance functions of S with mean zero at P_0. Moreover, since the operation of conditional expectation given S is projection onto this space, in a Type I or II problem we have

$$U_{CLR} = U - \Pi_{CLR}U = U - E(U \mid S) = U_c.$$

We note that U_{CLR} will depend on λ_0 unless we are in the Type I problem.

Since Λ_{CLR} is generally infinite dimensional, this representation gives us a problem no easier than finding the conditional expectation. To develop an approximation scheme, we begin with a nested sequence of finite dimensional subspaces of Λ_{CLR}, say $\Lambda_1 \subset \ldots \subset \Lambda_{CLR}$. If we let Π_t be the $\mathcal{L}_2$ projection operation onto Λ_t, then

$$U_t = U - \Pi_t U \tag{2}$$

is the t-th order projected score, and we might hope that $U_t \to U_{CLR}$ for an appropriately chosen sequence of subspaces. In this paper we will show that there is a natural stochastic basis, the Bhattacharyya scores, that generates a sequence of subspaces that is particularly advantageous. If we let $\Delta_t = U_t - U_{t-1}$ be the change in the projected scores, we will show in §5 that $n^{-1}\Delta_t$ is of stochastic magnitude $n^{-t/2}$, so that the changes are becoming progressively smaller.

The preceding stochastic results are independent of the type of model. If we wish to consider the use of projected scores to approximate a Type I conditional score, there remains the problem that the projected score $U_t(\psi, \lambda)$, although converging to a function of ψ alone, may still depend on λ. We will show in §6 that one can use $U_t(\psi, \hat{\lambda}_\psi)$, where $\hat{\lambda}_\psi$ is the maximum likelihood estimator of λ with ψ fixed, to approximate U_c without any increase in the stochastic order of magnitude of the error.

An early application of projected scores can be found in Neyman's $C(\alpha)$ test (Neyman, 1959). Lehmann (1981) provides a strong conceptual foundation for their development, and Small & McLeish (1988, 1989) present an extensive treatment. In addition to the above described stochastic properties, this paper is novel in the development of the multivariate parameter of interest. The fact that the extension is straightforward is important, as it contrasts with the difficulties faced by Cox & Reid (1987) in extension beyond the univariate case. In the final section we will offer some comments on future applications.

Throughout this paper expectations will always be taken at the model unless otherwise stated. The usual regularity conditions needed to take derivatives through integrals will be assumed and discussed where appropriate.

2 The Bhattacharyya basis and projected scores

2·1. The Bhattacharyya basis

Our goal of generality competes somewhat with our goal of clarity, as we wish to specify results in arbitrary finite dimension, but find that the notation sometimes obscures the simplicity. Therefore we will at times elucidate points only with the univariate case. Also some of the more technical results will be given in the appendix.

Given n independent observations with joint density $L(x;\psi,\lambda) = \prod f_k(x_k;\psi,\lambda)$ and log-likelihood $l(\psi,\lambda;x)$, denote the usual ψ-score $\nabla_\psi l(\psi,\lambda;x)$ by $U_0 = U_0(\psi,\lambda;x) = (U_0^1,\ldots,U_0^r)^T$, where

$$U_0^j = \frac{\partial\, L(x;\psi,\lambda)}{\partial\psi_j}\bigg/ L(x;\psi,\lambda), \qquad j = 1,\ldots,r.$$

Since the observations are independent U_0^j can be written as the sum, $\sum_{k=1}^n U_0^{jk}$ where $U_0^{jk} = \{\partial f_k(x_k;\psi,\lambda)/\partial\psi_j\}/f_k(x_k;\psi,\lambda)$.

Define $V_{i_1\ldots i_t}$ for $1 \le i_j \le s$ as

$$V_{i_1\ldots i_t} = \frac{\partial^t L(x;\psi,\lambda)}{\partial\lambda_{i_1}\ldots\partial\lambda_{i_t}}\bigg/ L(x;\psi,\lambda) \quad \text{and} \quad V^k_{i_1\ldots i_t} = \frac{\partial^t f_k(x_k;\psi,\lambda)}{\partial\lambda_{i_1}\ldots\partial\lambda_{i_t}}\bigg/ f_k(x_k;\psi,\lambda).$$

A fundamental property of the $V_{i_1\ldots i_t}$ is that under regularity they are all mean zero functions.

Define $\mathcal{V}_0$ to be the constant function 1 and let $\mathcal{V}_t$ denote the column vector containing all the t^{th} order derivative ratios $V_{i_1\ldots i_t}$. Finally define $\mathcal{V} = (\mathcal{V}_0^T,\ldots,\mathcal{V}_t^T)^T$. We note here that all functions are to be evaluated at the fixed parameter value of $\psi = \psi_0$. For the remainder of the paper it will be implicit that functions are evaluated at ψ_0 unless otherwise specified.

To illustrate the use of the V notation consider a situation in which there are 2 nuisance parameters, $s = 2$, and in which derivatives of up to the second order are taken so that $t = 2$. Then

$$\mathcal{V}_0 = 1, \qquad \mathcal{V}_1^T = (V_1, V_2), \qquad \mathcal{V}_2^T = (V_{11}, V_{12}, V_{21}, V_{22}),$$

and

$$\mathcal{V}^T = (1, \mathcal{V}_1^T, \mathcal{V}_2^T) = (1, V_1, V_2, V_{11}, V_{12}, V_{21}, V_{22}).$$

For a single nuisance parameter, $s = 1$, we would simply have $\mathcal{V}^T = (1, V_1, V_{11})$.

We refer to the set of functions $\{V_{i_1\ldots i_k}; 1 \le i_j \le s, j = 1,\ldots,k,\ k = 0,\ldots,t\}$ as the Bhattacharyya basis of order t in recognition of A. Bhattacharyya, who made extensive use of them in his study of information and unbiased estimation (Bhattacharyya, 1946). Let P_0 be the probability measure determined by the parameter values (ψ_0, λ_0) and denote the space of square integrable functions by $\mathcal{L}_2(P_0)$. Denote the r-fold Cartesian product of $\mathcal{L}_2(P_0)$ by $\mathcal{L}_2^r(P_0)$. Let Λ_t be the finite dimensional linear subspace of $\mathcal{L}_2(P_0)$ generated by all linear combinations of elements of the Bhattacharyya basis of order t. Its relationship with Λ_{CLR} will be found in a following subsection. We must consider the vector valued functions in the subspace Λ_t^r of $\mathcal{L}_2^r$ because ψ is $r-$dimensional. Here Λ_t^r denotes the r-fold Cartesian product of Λ_t.

For an arbitrary $\mathcal{L}_2^r$ function $h(x)$, let $\Pi_t h$ be the projection of h onto Λ_t^r. This projection is well defined since $\mathcal{L}_2^r$ is a Hilbert space and Λ_t^r is a closed subspace. For U_t as defined in (2), it can be shown that the projected score is

$$U_t = U_0 - E(U_0\,\mathcal{V}^T)\,E(\mathcal{V}\,\mathcal{V}^T)^{-}\mathcal{V}.$$

We note that U_1 is the score function for the least favorable family distribution and that U_2 is in the terminology of McLeish & Small (1988, p.68) second order unbiased.

For example, with ψ and λ univariate and $t = 2$, U_2 is equal to $U_0 - b_1 V_1 - b_2 V_{11}$ where the b_i are regression coefficients that depend only on the parameters and not the data. Also note that the function $\mathcal{V}_0 = 1$ is redundant since U_0 and all the $V_{i_1 \ldots i_k}$ are orthogonal to it. The projected score will be the fundamental object of study for the remainder of the paper.

2·2. G-Information

A useful tool in dealing with dealing with estimating functions is the information matrix. Following Godambe (1976) in spirit, we will say an estimative kernel g of the data X and parameter θ is G_0 -unbiased if for all θ, $E_\theta\,\{g(\theta; X)\} = 0$. Define the G-information in a G_0-unbiased kernel g to be $I_g = E(\nabla_\psi g)\,E(g\,g^T)^{-}\,E(\nabla_\psi g)^T$. Since g lies in $\mathcal{L}_2^r$, $\nabla_\psi g$ is an r by r matrix and likewise the information matrix I_g. The G-information corresponds to the inverse of the asymptotic variance of the root of g if g were used as an estimating function.

Let U_t^i denote the i^{th} component of the vector U_t. As a result of the regularity conditions, the information identity $E(U_t^i\,U_0^j) + E(\partial U_t^i/\partial\psi_j) = 0$, holds true. Using the orthogonality of U_t and $\mathcal{V}$ it follows that $E(U_t^i\, U_t^j) = -E(\partial U_t^i/\partial\psi_j)$. That is, $E(\nabla_\psi U_t) = -E(U_t\, U_t^T)$, thereby satisfying the "information identity" of Lindsay (1982). Substitution of this identity into the information definition results in:

$$\begin{aligned} J_t = I_{U_t} &= E(U_t\, U_t^T) \\ &= E(U_0\, U_0^T) - E(U_0\, \mathcal{V}^T)\, E(\mathcal{V}\, \mathcal{V}^T)^-\, E(\mathcal{V}\, U_0^T). \end{aligned} \tag{3}$$

Although these information matrices can be used to develop an optimality theory for score functions with various degrees of "debiassing" (Waterman, 1993), our use here relates to certain of their properties. One important set of properties arise due to the nesting of the subspaces Λ_t : we have $J_1 \geq J_2 \geq ...$, with

$$\mathrm{var}(U_a - U_b) = J_b - J_a \text{ for } b \geq a \tag{4}$$

Moreover, in the type I or II model, $J_t \geq I_{U_c}$, the conditional information. In the latter case, it can be shown that we can measure the stochastic discrepancy between U_c and U_t by evaluating the difference in their information matrices:

$$\mathrm{var}(U_t - U_c) = \mathrm{var}(U_t) - \mathrm{var}(U_c).$$

Some examples of this calculation will be given in the next section.

A useful tool for performing calculations regarding the projected scores is the extended information matrix M_t, where

$$M_t = E\left\{ \begin{pmatrix} U_0 \\ \mathcal{V} \end{pmatrix} \left(U_0^T, \mathcal{V}^T\right) \right\} = \begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix}.$$

In this notation $U_t = U_0 - I_{\psi\lambda}\, I_{\lambda\lambda}^- \mathcal{V}$ and $J_t = I_{\psi\psi} - I_{\psi\lambda}\, I_{\lambda\lambda}^-\, I_{\lambda\psi}$. For the case $t = 1$, M_1 is the usual Fisher's information matrix and, in the case $r = s = 1$,

$$M_2 = E \begin{pmatrix} U_0 U_0 & U_0 V_1 & U_0 V_{11} \\ V_1 U_0 & V_1 V_1 & V_1 V_{11} \\ V_{11} U_0 & V_{11} V_1 & V_{11} V_{11} \end{pmatrix}.$$

Finally, we note that the projected scores have ideal invariance properties. If the nuisance parameters are subject to a continuously differentiable one-to-one transformation, which may depend on the parameter of interest, it is easily checked that the projected scores are invariant under this class of interest preserving transformations.

2·3. *Convergence questions*

The question of whether $U_t \to U_{CLR}$ as $t \to \infty$ cannot be answered fully here, but we believe it useful to explain some of the mathematical issues involved. First, since by construction Λ_{CLR} is closed, functions that arise by taking differences between elements, renorming, and taking limits are included. Therefore, provided variances are finite, the Bhattacharyya basis functions are elements: $\Lambda_t \subset \Lambda_{CLR}$. Next, let $T_t(\lambda)$ be the t-term Taylor expansion in λ of the likelihood ratio function $C(\lambda) = \{f(x;\psi_0,\lambda)/f(x;\psi_0,\lambda_0)\} - 1$ about the point λ_0. Suppose that this Taylor expansion converges in $\mathcal{L}_2$, namely

$$E\{C(\lambda) - T_t(\lambda)\}^2 \to 0 \quad \text{as } t \to \infty. \tag{5}$$

Then, since each $T_t(\lambda)$ is in Λ_t, it is clear that $clos(\cup_t \Lambda_t) = \Lambda_{CLR}$.

The spanning condition (5) suffices to guarantee the desired convergence of scores. For simplicity, consider the scalar case. The declining nature of the score variances, together with the difference property (4), ensures that U_t is a Cauchy sequence in $\mathcal{L}_2$. If we let its limit be U_∞, then $U_\infty \perp \Lambda_t$ for all t, and therefore U_∞ is orthogonal to the basis elements $C(\lambda)$ of Λ_{CLR} for all λ :

$$E^2\{U_\infty C(\lambda)\} = E^2[U_\infty\{C(\lambda) - T_t(\lambda)\}] \leq E(U_\infty)^2 E\{C(\lambda) - T_t(\lambda)\}^2 \to 0.$$

But U_{CLR} is the unique projection satisfying this orthogonality criterion, so $U_\infty = U_{CLR}$.

Further analysis of convergence questions related to Bhattacharyya type projections are found in the literature on minimum variance unbiased estimation. See, for example, Ghosh & Sathe (1987) and Mase (1977) for results on exponential family models. See also Small & McLeish (1991, p. 271) for a further discussion.

2·4. *Other sampling schemes*

The case of univariate ψ and λ_i will be described, solely for ease of exposition, rather than technical necessity. The i superscript now refers to the ith stratum. Suppose that from a set of p strata (or subjects) we take independent observations, and that each strata has its own nuisance parameter λ_i, so that the density has the form

$$f(x;\psi,\lambda_1,\ldots,\lambda_p)=\prod_{i=1}^{p}\prod_{j=1}^{n_p} f(x_{ij};\psi,\lambda_i).$$

In this case a Bhattacharyya score with cross derivatives, such as $V_{i_1 i_2}$, depends on the data in more than one strata. However it is easily shown that these terms have zero weight in the projected score, so that the projected score is simply the sum of the separate projected scores in each stratum. The informations also add over the strata. That is

$$U_t=\sum_{i=1}^{p} U_t^i,\ \text{and}\ J_t=I_{U_t}=\sum_{i=1}^{p} E(U_t^i U_t^i).$$

Another important type of structure arises in regression problems, where the method again has straightforward application. For example, consider a generalized linear model with canonical link, known dispersion parameter and loglikelihood kernel

$$l(\theta;y)=\sum_{i=1}^{n} y_i\eta_i-\sum_{i=1}^{n} b_i(\eta_i) \quad \text{where } \theta=\begin{pmatrix}\psi\\ \lambda\end{pmatrix} \text{ and } \eta_i=x_i\theta.$$

Corresponding to the partition of θ there is a partition of x_i, say $x_i=(z_i,\gamma_i)$, so that in the obvious notation $X=(Z,\Gamma)$. This is a model of Type I structure, with $S=\Gamma^T y$. The exact conditional structure is potentially quite difficult, but the calculation of a second order projected score is straightforward. We have $U_0^j=\partial l(\theta;y)/\partial\psi_j=\sum y_i z_{ij}-\sum_i b'(\eta_i)z_{ij}$ where $'$ denotes differentiation with respect to η_i. Hence $U_0=Z^T\{Y-b'(\eta)\}$, where $b'(\eta)$ is a vector whose i^{th} element is $b'(\eta_i)$. By analogy $\mathcal{V}_1=\Gamma^T\{Y-b'(\eta)\}$, and using (6) it is simple to show that V_{ij} is the $(i,j)^{th}$ element of the matrix $\Gamma^T[\{Y-b'(\eta)\}\{Y-b'(\eta)\}^T-b''(\eta)]\,\Gamma$, where $b''(\eta)$ is a diagonal matrix with elements $b''(\eta_i)$. Waterman & Lindsay (1994) consider several examples of this type in which the second order projected score gives good approximations to the conditional score.

3 Examples

In this section we give a couple of simple but important examples that illustrate the methodology and show that the approximation of the projected score U_t to the conditional score can be quite accurate, even for t as small as 2.

Example 1. The canonical Neyman-Scott problem has n independent normally distributed observations from each of p independent strata. Every stratum has its own mean λ_i, and the parameter of interest is the common variance ψ. This is a Type I problem, with the conditioning statistics being the stratum sample means. Our earlier remarks justify doing the projection calculations for each stratum individually. The following calculations are carried out for the i^{th} stratum, the i index referring to the stratum:

$$\begin{aligned}\frac{\partial f(x_i;\psi,\lambda_i)}{\partial\psi} = U_0^i &= -\frac{n}{2\,\psi} + \frac{1}{2\,\psi_2}\sum_{j=1}^{n}(x_{ij}-\lambda_i)^2 \\ \frac{\partial f(x_i;\psi,\lambda_i)}{\partial\lambda_i} = V_i &= \frac{1}{\psi}\sum_{j=1}^{n}(x_{ij}-\lambda_i)\end{aligned}$$

The M_2 extended information matrix is

$$M_2 = \begin{pmatrix} \frac{n}{2\psi^2} & 0 & \frac{n}{\psi^2} \\ 0 & \frac{n}{\psi} & 0 \\ \frac{n}{\psi^2} & 0 & \frac{2n^2}{\psi^2} \end{pmatrix},$$

and $U_2^i = U_0^i - \frac{1}{2n}V_{ii}$. A straightforward calculation shows that $U_2 = U_c$ in this problem. This example appears in McLeish & Small (1988, p.68) but the link to the conditional score function is not made.

Example 2. Another classic model with conditional structure consists of two independent binomials X and Y parameterized as follows:

$$X \sim Bin\left(m_1, \frac{exp(\psi+\lambda)}{1+exp(\psi+\lambda)}\right) \quad \text{and} \quad Y \sim Bin\left(m_2, \frac{exp(\lambda)}{1+exp(\lambda)}\right).$$

The log odds ratio ψ is the parameter of interest, and a complete sufficient statistic for the nuisance parameter λ is $X+Y$. Since this is a Type I problem an optimal estimating equation exists for ψ, being the conditional score U_c, which has the computationally tedious form:

$$U_c = X - \frac{\displaystyle\sum_{i=max(0,t-m_2)}^{min(m_1,t)} \binom{m_1}{i}\binom{m_2}{t-i} i\, exp(\psi\, i)}{\displaystyle\sum_{i=max(0,t-m_2)}^{min(m_1,t)} \binom{m_1}{i}\binom{m_2}{t-i} exp(\psi\, i)} \quad \text{and} \quad t = x+y.$$

Routine calculations yield, with $'$ denoting differentiation with respect to λ:

$$\begin{aligned}
U_0 &= -m_1\frac{exp(\psi+\lambda)}{1+exp(\psi+\lambda)}+x \\
V_1 &= -m_1\frac{exp(\psi+\lambda)}{1+exp(\psi+\lambda)}+x-m_2\frac{exp(\lambda)}{1+exp(\lambda)}+y \\
V_1' &= m_1\frac{exp(2\lambda+2\psi)}{\{1+exp(\lambda+\psi)\}^2}-m_1\frac{exp(\lambda+\psi)}{1+exp(\lambda+\psi)}+m_2\frac{exp(2\lambda)}{\{1+exp(\lambda)\}^2}-m_2\frac{exp(\lambda)}{1+exp(\lambda)} \\
V_{11} &= V_1'+V_1^2
\end{aligned}$$

Note that V_1 is linear in the sufficient statistic $X+Y$ and that V_{11} is quadratic in $X+Y$. The relevance of this observation will be expanded on in §6.

The U_1 and U_2 informations were calculated, and compared to the information in the conditional score function, $E(U_c^2) = -E(\partial U_c/\partial\psi)$. Table 1 shows the various informations when $\psi = 0.5$ and $\lambda = -0.25$. These parameter values correspond to probabilities of success in the two binomials 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 presented, this transformation being appropriate for the scale of the parameter estimates. The sample sizes, m_1 and m_2 have been set equal, consequently the experiment is equivalent to observing m_1 independent Bernoulli pairs (X_i, Y_i) so that the U_1 information increases linearly in m_1.

TABLE 1 ABOUT HERE

Note that $U_2 = U_c$ for the case when $m_1 = m_2 = 1$, that is the Bernoulli problem, or binary matched pairs. It follows that as an approximation to the conditional score, in this situation, the projected methodology improves on the "adjusted profile log-likelihood" and the "modified profile log-likelihood" (Barndorff-Nielsen, 1986), illustrated in McCullagh & Tibshirani (1990, p.337).

It is immediately apparent from the table 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. (Recall also that the difference in informations equals the variance of the difference between score functions).

4 Some basic properties of the Bhattacharyya basis

We now present two identities that are repeatedly used to simplify asymptotic expansions involving the Bhattacharyya basis functions. The first of these identities provides a recurrence formula for expressing higher order Bhattacharyya basis elements in terms of the derivatives and products of lower order elements:

$$V_{i_1 \ldots i_{t+1}} = \frac{\partial V_{i_1 \ldots i_t}}{\partial \lambda_{i_{t+1}}} + V_{i_1 \ldots i_t} V_{i_{t+1}}. \tag{6}$$

As an example of the use of the formula, first consider the simplest case, that is when there is a single nuisance parameter λ, then (6) becomes $V_{11} = \partial V_1 / \partial \lambda_1 + V_1 V_1$. In the case of two nuisance parameters, $\lambda = (\lambda_1, \lambda_2)$ then one of the identities that (6) yields is $V_{12} = \partial V_1 / \partial \lambda_2 + V_1 V_2$. The recurrence formula proves very useful when calculating the Bhattacharyya functions of high order through the use of a symbolic computational software package.

The use of a partially orthogonalized version of the Bhattacharyya basis leads to simplifications in proofs that involve the projected scores. In particular, a partially orthogonalized basis will be used in which the Bhattacharyya functions of order t are made orthogonal to all lower order Bhattacharyya functions, but are allowed to be correlated among themselves. These functions will be denoted by $\mathcal{V}_t^*$. Using the orthogonalized functions gives rise to an extended information matrix of the form

$$M_t = \begin{pmatrix} E(U_0 U_0^T) & E(U_0 \mathcal{V}_1^T) & E(U_0 \mathcal{V}_2^{*T}) & E(U_0 \mathcal{V}_2^{*T}) & \cdots & E(U_0 \mathcal{V}_t^{*T}) \\ E(\mathcal{V}_1 U_0^T) & E(\mathcal{V}_1 \mathcal{V}_1^T) & 0 & 0 & \cdots & 0 \\ E(\mathcal{V}_2^* U_0^T) & 0 & E(\mathcal{V}_2^* \mathcal{V}_2^{*T}) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ E(\mathcal{V}_t^* U_0^t) & 0 & 0 & \cdots & \cdots & E(\mathcal{V}_t^* \mathcal{V}_t^{*T}) \end{pmatrix}.$$

As a convention $\mathcal{V}_1^*$ will be set equal to $\mathcal{V}_1$.

In the process of performing the calculations involving these orthogonalized functions, it was found that important simplifications arose because of certain higher order analogues of the information identities. We give the necessary technical results first, then illustrate their application. For the remainder of §4 the Bhattacharyya basis functions are evaluated at λ in a neighborhood of λ_0. Orthogonality statements are with respect to expectations taken under $P(\lambda)$ and are assumed to be true for all λ in a neighborhood of λ_0.

Let $T(\lambda) \in \mathcal{G}_0$ be a smooth function of λ that is orthogonal to the Bhattacharyya basis of order $t-1$. Consider a partition of the indices $\{i_1, \ldots, i_t\}$ into two disjoint sets υ_1 and υ_2. Denote the derivative of T with respect to $\lambda_{i_1}, \ldots, \lambda_{i_q}$ by $T_{i_1 \ldots i_q}$ where q is the number of indices in υ_1. As usual the function $\mathcal{V}_0$ is the constant function 1.

Proposition 4.1: If $T(\lambda) \perp \{\mathcal{V}_0, \ldots, \mathcal{V}_{t-1}\}$ *then*

$$E(T_{\upsilon_1} V_{\upsilon_2}) = (-1)^{t-q} E(T_{i_1 \ldots i_t}) \qquad 0 \leq q \leq t.$$

From a symbolic perspective, the proposition states that the subscripts can be exchanged between the T and V functions, with the sign of the result alternating between -1 and $+1$ according to whether an odd or even number of subscripts have been exchanged. The proof will be found in the appendix.

Corollary 4.1: If $T(\lambda) \perp \{\mathcal{V}_0, \ldots, \mathcal{V}_{t-1}\}$ *then*

$$E(T_{i_1 \ldots i_p}) = 0, \text{ for } p = 0, \ldots, t-1.$$

Proof: If $T \perp \{\mathcal{V}_0, \ldots, \mathcal{V}_{t-1}\}$, then $T \perp \{\mathcal{V}_0, \ldots, \mathcal{V}_{p-1}\}$ for any $p \leq t-1$. Applying Proposition 4.1 with $t = p$ and $q = 0$ implies that

$$E(T V_{i_1 \ldots i_p}) = (-1)^p E(T_{i_1 \ldots i_p}),$$

but since $p \leq t-1$ and $T \perp \{\mathcal{V}_0, \ldots, \mathcal{V}_{t-1}\}$, $E(T V_{i_1 \ldots i_p})$ must equal 0 which gives the result. □

To illustrate the application of these results in the simplest possible setting consider the case where the nuisance parameter λ is univariate. The function V_1 is orthogonal to the constant function, so that the proposition applies to it with $t = 1$; setting $q = 0$ gives $E(V_1 V_1) = -E(V_1')$, the usual information identity. An interesting extension is for the function V_{11}^* which is orthogonal to both the constant function and V_1 so that Proposition 4.1 applies to it with $t = 2$. Setting $q = 0$ results in the identity $E(V_{11}^* V_{11}^*) = E(V_{11}^*{}'')$, where $'$ denotes differentiation with respect to λ. This last example, which is the first in a series of "extended information identities", shows some of the usefulness of the results, since an expectation involving derivatives of Bhattacharyya basis elements has been expressed as a variance of a basis element.

Finally, we note that since $U_t \perp \{\mathcal{V}_0, \ldots, \mathcal{V}_t\}$, the proposition and corollary apply to its derivatives. In particular, all its derivatives of order up to t have mean zero. Thus, for example, in a Taylor series expansion in λ, a term such as U_2' or U_2'' has mean zero, a fact of clear potential import in asymptotic properties. Moreover $E(U_t^{(t+1)}) = (-1)^{t+1} E(U_t V_{i_1 \ldots i_{t+1}})$, so that in particular $E(U_1'') = E(U_1 V_{11}) = E(U_1 V_{11}^*) = E(U_0 V_{11}^*)$. The fact that $E(U_1'') \neq 0$ but that $E(U_2'') = 0$ will turn out to be important in asymptotic analyses.

5 Asymptotic and constructive properties

We now demonstrate the theoretical strength of the Bhattacharyya basis through an asymptotic analysis. Write the U_t score as

$$U_t = U_1 + \sum_{j=2}^{t} \Delta_j,$$

where the $\Delta_j = U_j - U_{j-1}$. The Δ_j are uncorrelated and the following theorem identifies their stochastic order of magnitude where all expectations and functions are evaluated at (ψ_0, λ_0).

Theorem 5.1:

$$I_{U_1} = \mathrm{var}(U_1) = n(i_{\psi\psi} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\psi}), \qquad tr\{\mathrm{var}(U_1)\} = O(n),$$

where $i_{\psi\psi}, i_{\lambda\lambda}, i_{\psi\lambda}$ and $i_{\lambda\psi}$ are the elements of the usual Fishers Information matrix based on a sample of size 1. For $\Delta_t = U_t - U_{t-1}, \qquad t = 1, 2, \ldots$

$$\mathrm{tr}\{\mathrm{var}(\Delta_t)\} = \mathrm{tr}\{\mathrm{var}(U_{t-1}) - \mathrm{var}(U_t)\} = \mathrm{tr}(I_{U_{t-1}} - I_{U_t}) = O(n^{2-t}).$$

The proof of this theorem requires the development of techniques for evaluating the magnitude of the entries of the extended information matrix M_t. This, in turn, involves a detailed understanding of the Bhattacharyya scores. The key to our understanding of their properties are their representations in U-statistic form. For example, for univariate λ, the Bhattacharyya basis functions V_1, V_{11} and V_{111} have representations:

$$V_1 = \sum_{r_1=1}^{n} V_1^{r_1}, \quad V_{11} = \sum_{r_1=1}^{n} V_{11}^{r_1} + \sum_{r_1 \neq r_2}^{n} V_1^{r_1} V_1^{r_2}$$

$$V_{111} = \sum_{r_1=1}^{n} V_{111}^{r_1} + 3 \sum_{r_1 \neq r_2}^{n} V_{11}^{r_1} V_1^{r_2} + \sum_{r_1 \neq r_2 \neq r_3}^{n} V_1^{r_1} V_1^{r_2} V_1^{r_3}$$

and for the simplest multivariate case with $\dim(\lambda) = 2$,

$$V_{12} = \sum_{r_1=1}^{n} V_{12}^{r_1} + \sum_{r_1 \neq r_2}^{n} V_1^{r_1} V_2^{r_2}.$$

In each case, each summed term in the representation is uncorrelated with every other such term, so that the variances add and magnitudes in n are easily determined. Covariances between the basis elements are similarly easy to determine, thereby giving the order of magnitude of the elements of M_t.

We can summarize the key points in the following representation result:

Proposition 5.1: The t-th order Bhattacharyya basis element can be written as

$$V_{i_1 i_2 \ldots i_t} = \sum_{1 \leq \alpha \leq t} h(\alpha).$$

where each $h(\alpha)$, *when divided by* $\binom{n}{\alpha}$ *is a mean zero U-statistic of order* α.

For $\alpha_1 \neq \alpha_2$, $E\{h(\alpha_1) h(\alpha_2)\} = 0$.

Moreover,

$$E(V_{i_1 \ldots i_{t_1}} V_{j_1 \ldots j_{t_2}}) = O(n^{min(t_1, t_2)}), \tag{7}$$

a special case of which is $\mathrm{var}(V_{i_1 \ldots i_{t_1}}) = O(n^{t_1})$.

Some further details on these representations are in the appendix.

The usual likelihood theory relies heavily on representations of the score function as an independently identically distributed sum so that the Law of Large Numbers and Central Limit Theorem apply. Higher order asymptotics built around the Bhattacharyya basis functions will also be tractable, because the basis functions are comprised of the simplest extension of the independently identically distributed sum; they have representations as sums of uncorrelated U-Statistics.

Finally, we turn to the proof of Theorem 5.1. Here we just outline the proof in the univariate λ case and point out the technical details which can be found in (Waterman, 1993, Ch. 3).

First note that $\mathrm{var}(\Delta_t) = E(U_0 \mathcal{V}_t^{*T}) E(\mathcal{V}_t^* \mathcal{V}_t^*)^- E(\mathcal{V}_t^* U_0^T)$. Assume that $E(U_0 \mathcal{V}_t^*)$ has the same order of magnitude as $E(U_0 \mathcal{V}_t)$, that $E(\mathcal{V}_t^* \mathcal{V}_t^*)$ has the same order of magnitude as $E(\mathcal{V}_t \mathcal{V}_t)$ and that $E(\mathcal{V}_t^* \mathcal{V}_t^*)^-$ has order of magnitude equal to the inverse of the order of magnitude of $E(\mathcal{V}_t^* \mathcal{V}_t^*)$. Since U_0 is an independently identically distributed sum it follows that $E(U_0 \mathcal{V}_t) = O(n)$, and (7) states that $E(\mathcal{V}_t \mathcal{V}_t) = O(n^t)$. Consequently $\mathrm{var}(\Delta_t) = O(n) O(n^{-t}) O(n) = O(n^{2-t})$ as required. □

The technical details involve establishing that $E(U_0 \mathcal{V}_t^*)$ really does have the same order of magnitude as $E(U_0 \mathcal{V}_t)$ and that the order of magnitude of $E(\mathcal{V}_t^* \mathcal{V}_t^*)^-$ is $O(n^{-t})$. This last fact can be shown to follow from the continuity of the generalized inverse for a class of matrices that includes $E(\mathcal{V}_t^* \mathcal{V}_t^*)$.

6 Properties of U_t and substitution approximations $\hat{U}_t$ to U_c in the exponential family

We now focus on the use of the projected scores in the Type I model, where there exists a complete sufficient statistic for the nuisance parameter not depending on ψ, so that there exists a conditional score function $U_c(\psi)$ that is free of λ. The key result of this section demonstrates that in a standard exponential family setting, the substitution version of the projected score $U_t(\psi, \hat{\lambda}_\psi)$ is remarkably close to the conditional score.

The difference of interest to us, $U_t(\psi, \hat{\lambda}_\psi) - U_c(\psi)$, can be expressed as

$$\{U_t(\psi, \hat{\lambda}_\psi) - U_t(\psi, \lambda_0)\} + \{U_t(\psi, \lambda_0) - U_c(\psi)\}.$$

Theorem 5.1 suggests that the second term $U_t(\psi, \lambda_0) - U_c(\psi) = \sum_{i=t+1}^{\infty} \Delta_i$ has variance of order $O(n^{1-t})$, as this is the variance of the largest stochastic term in the right hand sum. Theorem 6.1 will show that the first term $U_t(\psi, \hat{\lambda}_\psi) - U_t(\psi, \lambda_0)$ also has asymptotic variance of the same order of magnitude, that is $O(n^{1-t})$. The results are presented for univariate ψ and λ but the obvious multivariate extensions hold.

Since λ is taken to be univariate we will use a condensed notation for the Bhattacharyya basis elements with V_t being $(\partial^t L/\partial\lambda^t)/L$ and $\mathcal{V}_t$ being the entire Bhattacharyya basis of order t. Finally all derivatives are taken with respect to λ and are denoted by $'$. We first detail further properties of the Bhattacharyya scores in a standard exponential family setting, returning to our main result at the end.

The results will be for a sample of size n from the following canonically parameterized exponential family

$$f(x_i; \psi, \lambda) = \exp\{\psi x_i + \lambda S(x_i) - \tau(\psi, \lambda)\} \tag{8}$$

for which a complete sufficient statistic for λ with ψ fixed is $S = \sum S(x_i)$. The following proposition identifies some important algebraic simplifications that occur in this setting:

Proposition 6.1: In model (8), V_t is a monic polynomial in S of degree t, and both V_t' and $V_t^{*\prime}$ are polynomials in S of degree $t-1$. Furthermore, writing U_t as $U_0 - \sum_{i=1}^t a_i V_i^*$ we have the surprising identity:

$$U_t' = -a_t' V_t^*, \qquad t = 0, 1 \ldots$$

and as a consequence, the stochastic magnitude of $U_t^{(k)}$ is $O_p(n^{-\frac{1}{2}(t-k-1)})$.

Proof: Differentiating the log-likelihood with respect to λ establishes V_1 as linear in S. The fact that V_t is a monic polynomial in S of degree t can be shown by induction using the recurrence relation given in (6). Since both V_t and V_t^* are polynomials in S of degree t with the coefficient of S^t being 1, it follows that both V_t' and $V_t^{*\prime}$ are polynomials in S of degree $t-1$.

To show the identity $U_t' = -a_t' V_t^*$ for the cases $t = 0$ and $t = 1$ direct evaluation of U_0' and U_1' gives the result. For the general case, $t \geq 2$, by appealing to Corollary 4.1 it can be shown that $U_t' \perp \mathcal{V}_{t-1}$ and $V_t^{*\prime} \perp \mathcal{V}_{t-2}$, $t = 2, 3, \ldots$. Since $V_t^{*\prime}$ lies in the space spanned by $\mathcal{V}_{t-1}$, the space of polynomials of degree $t-1$, and is orthogonal to $\mathcal{V}_{t-2}$ it then follows that $V_t^{*\prime} = c_t V_{t-1}^*$, for some appropriate constant c_t. By construction

$$U_t = U_0 - \sum_{i=1}^{t} a_i V_t^*,$$

where a_i are the constants that make U_t orthogonal to $\mathcal{V}_t$. Therefore

$$\begin{aligned} U_t' &= U_0' - \sum_{i=1}^{t} a_i' V_i^* - \sum_{i=1}^{t} a_i V_i^{*\prime} \\ &= U_0' - \sum_{i=1}^{t} a_i' V_i^* - \sum_{i=1}^{t} a_i c_i V_{i-1}^* \\ &= U_0' - \sum_{i=1}^{t-1} b_i V_i^* - a_1 c_1 V_0^* - a_t' V_t^* \end{aligned}$$

where $b_i = a_i' + a_{i+1} c_{i+1}$. For model (8) U_0' is a constant, and using the fact that $U_t' \perp \mathcal{V}_{t-1}$ and by taking the inner product of U_t' with $1, V_1^*, \cdots, V_{t-1}^*$ it follows that $b_1 = b_2 = \cdots = b_{t-1} = 0$. Hence $U_t' = -a_t' V_t^*$ as required. The arguments leading to the statement that $U_t^{(k)}$ is $O_p(n^{-\frac{1}{2}(t-k-1)})$ can be found in the proof of Theorem 6.1 in the appendix. □

Since in the Type I exponential family problem the conditional score function is completely free of the nuisance parameter λ it is immediate that $\partial U_c / \partial \lambda \equiv 0$. The above proposition confirms that $\partial U_t / \partial \lambda$ has only a weak dependence on λ in a Type I model. This, in turn, plays a critical role in a Taylor series proof of the following result.

Theorem 6.1: In the Type I model described above

$$U_t(\hat{\lambda}_\psi) - U_t(\lambda_0) = O_p(n^{-\frac{1}{2}(t-1)}).$$

Proof: The key to the proof is the following Taylor series which expands the function $n^{-\frac{1}{2}(t-1)}\{U_t(\hat{\lambda}_\psi) - U_t(\lambda_0)\}$ about λ_0:

$$\begin{aligned} n^{-\frac{1}{2}(t-1)}(U_t(\hat{\lambda}_\psi) - U_t(\lambda_0)) &= n^{\frac{1}{2}(t-1)} \sum_{k=1}^{t+1} \frac{1}{k!} (\hat{\lambda}_\psi - \lambda_0)^k U_t^{(k)}(\lambda_0) + \\ & \quad n^{\frac{1}{2}(t-1)} \frac{1}{(t+2)!} (\hat{\lambda}_\psi - \lambda_0)^{t+2} U_t^{(t+2)}(\lambda^*) \qquad (9) \end{aligned}$$

where $|\hat{\lambda}_\psi - \lambda_0| > |\lambda^* - \lambda|$. Our aim is to show that all terms in the Taylor series are $O_p(1)$ and that the remainder is $o_p(1)$. The important point is that the result of Theorem 6.1 equates $U_t^{(k)}$ with $(-a_t' V_t^*)^{(k-1)}$ from which point the proof amounts to order of magnitude bookkeeping.

Further discussion of the proof is in the appendix. □

7 Concluding remarks

This work is the first part of a continuing investigation of the properties of projected scores. In this paper, we have largely confined ourselves to the manner in which these scores can be used to approximate conditional scores, of either Type I or II, by creating a finite dimensional, and hence algebraically straightforward, approximation method that has the bonus of an extreme accuracy that can be partially explained by its asymptotic properties. In the Type I problem, we have shown that by using the maximum likelihood estimator of the nuisance parameter, we can approximate to arbitrary stochastic accuracy the conditional score function.

There are a large number of topics which are subjects of our further investigation. For the Type II problem, in which U_c depends on λ, there is a natural corresponding problem of showing that $U_t(\hat{\lambda}_\psi)$ is close to $U_c(\hat{\lambda}_\psi)$. While we believe this to be generally true, we do not prove it here, as it is rightfully part of a fuller investigation of the properties of the projected scores in Type II or III models; in those structures, the nuisance parameter cannot be fully eliminated, and so stochastic properties depend on the method of estimation of this parameter. Among our results remaining to be reported are the role of statistical curvature in determining the structure of the projected scores and their consequent asymptotic virtues for removing score bias in a Neyman Scott type framework.

Appendix

Proof of proposition 4.1

In this proof derivatives of f with respect to $\lambda_{i_1}, \ldots, \lambda_{i_r}$ are denoted by $f_{i_1 \ldots i_r}$.

Proof: The proof is by induction on t. The case when $t = 1$ follows from the assumption that the derivative $\partial/\partial\lambda_{i_1}$ can be taken through the integral $\int T\,f dx = 0$. The first interesting case is $t = 2$ when the theorem asserts the following three identities: $E(T_{i_1 i_2} V_0) = E(T_{i_1 i_2})$, $E(T_{i_1} V_{i_2}) = -E(T_{i_1 i_2})$ and $E(T V_{i_1 i_2}) = E(T_{i_1 i_2})$.

The first of these is trivially true. The second is obtained from differentiating the identity $\int T_{i_1} f dx = 0$ once with respect to λ_{i_2}, and the third arises from differentiating the identity $\int T f dx = 0$ twice with respect to λ_{i_1} and λ_{i_2} and using the fact that $E(T_{i_1} V_{i_2}) = -E(T_{i_1 i_2})$.

Now assume that the result is true for $T \perp \{\mathcal{V}_0, \ldots \mathcal{V}_{t-2}\}$. Then

$$\begin{aligned} E(T_{i_1 \ldots i_q} V_{i_{q+1} i_{q+2} \ldots i_t}) &= \int T_{i_1 \ldots i_q} f_{i_{q+1} \ldots i_t} dx \\ &= \int \left(T_{i_1 \ldots i_q} f_{i_{q+1} \ldots i_t} + T_{i_1 \ldots i_{q+1}} f_{i_{q+2} \ldots i_t} - T_{i_1 \ldots i_{q+1}} f_{i_{q+2} \ldots i_t} \right) dx \\ &= \int \partial(T_{i_1 \ldots i_q} f_{i_{q+2} \ldots i_t}) \big/ \partial\lambda_{i_{q+1}} dx - \int T_{i_1 \ldots i_{q+1}} f_{i_{q+2} \ldots i_t} dx \\ &= \partial E(T_{i_1 \ldots i_q} V_{i_{q+2} \ldots i_t}) \big/ \partial\lambda_{i_{q+1}} - E(T_{i_1 \ldots i_{q+1}} V_{i_{q+2} \ldots i_t}). \end{aligned}$$

But by the induction hypothesis

$$E(T_{i_1 \ldots i_q} V_{i_{q+2} \ldots i_t}) = (-1)^{t-(q+1)} E(T_{i_1 \ldots i_q i_{q+2} \ldots i_t}) = (-1)^q E(T V_{i_1 \ldots i_q i_{q+2} \ldots i_t}) = 0$$

from the orthogonality assumption. It follows that

$$E(T_{i_1 \ldots i_q} V_{i_{q+1} \ldots i_t}) = -E(T_{i_1 \ldots i_{q+1}} V_{i_{q+2} \ldots i_t}).$$

The same argument is repeated until all the subscripts have been transferred from V to T, which provides the result. □

Proof of Proposition 5.1

The theorems presented are valid for independently identically distributed samples. The first task is to express $V_{i_1 \ldots i_t}$ in a form that can be used to recover the U-statistic representation. The key operation in forming $V_{i_1 \dots i_t}$ is to obtain the high order derivative $\partial^t/(\partial\lambda_{i_1}\dots\partial\lambda_{i_t})$ of the composite function $\prod f_k(x_k)$. Establishing a notation for such an operation amounts to establishing a notation for partitions of the set of indices $\{i_1, i_2, \dots i_t\}$, a task that is conveniently handled by notation from the theory of lattices. The partial order relation on the partitions is that $a < b$ if a is a sub-partition of b. The reader is referred to McCullagh (1987) for an extensive presentation of the notation. An arbitrary partition of t indices $\{i_1, i_2, \dots, i_t\}$ into α blocks will be denoted by $\Upsilon = \{\upsilon_1, \upsilon_2, \dots, \upsilon_\alpha\}$, where each υ_i contains a subset of the t indices. The largest partition will be denoted by Υ_1 and the smallest by Υ_t, so that Υ_t corresponds to the partition of the indices into t blocks. Using this notation then $V_{i_1 \dots i_t}$ can be expressed as:

$$V_{i_1 i_2 \dots i_t} = \sum_{\Upsilon_t \le \Upsilon \le \Upsilon_1} g(\Upsilon) \quad \text{where} \quad g(\Upsilon) = \sum_{r_1 \ne r_2 \dots \ne r_\alpha} V_{\upsilon_1}^{r_1} V_{\upsilon_2}^{r_2} \dots V_{\upsilon_\alpha}^{r_\alpha} [\tau(\Upsilon)],$$

and $\tau(\Upsilon)$ is the number of non-distinct partitions of Υ, that is, partitions that are equivalent up to identification of the indices.

The bracket [ ] notation is used in order to avoid explicitly listing the non-distinct partitions (McCullagh, 1987, p.29).

To clarify the notation consider the case where $t = 3$ so that $V_{i_1 i_2 i_3}$ is of interest. Then the possible partitions of $\{i_1 i_2 i_3\}$ are $\Upsilon_1 = \{i_1 i_2 i_3\}$, $\{i_1 i_2 | i_3\}$ and $\Upsilon_3 = \{i_1 | i_2 | i_3\}$. It follows that

$$\begin{aligned} V_{i_1 i_2 i_3} = \sum_{r_1} V_{i_1 i_2 i_3}^{r_1} [\tau(\{i_1 i_2 i_3\})] \quad &+ \sum_{r_1 \ne r_2} V_{i_1 i_2}^{r_1} V_{i_3}^{r_2} [\tau(\{i_1 i_2 | i_3\})] \\ &+ \sum_{r_1 \ne r_2 \ne r_3} V_{i_1}^{r_1} V_{i_2}^{r_2} V_{i_3}^{r_3} [\tau(\{i_1 | i_2 | i_3\})]. \end{aligned}$$

In this case

$$[\tau(\{i_1 i_2 i_3\})] = [1], \quad [\tau(\{i_1 i_2 | i_3\})] = [3], \quad [\tau(\{i_1 | i_2 | i_3\})] = [1],$$

since there is only one combination of the indices $\{i_1 i_2 i_3\}$ but three of the indices $\{i_1 i_2 | i_3\}$ namely $\{i_1 i_2 | i_3\}$,$\{i_1 i_3 | i_2\}$ and $\{i_2 i_3 | i_1\}$, and again there is only one combination of the indices $\{i_1|i_2|i_3\}$. To be precise the term

$$\sum_{r_1 \neq r_2} V^{r_1}_{i_1 i_2} V^{r_2}_{i_3} [\tau(\{i_1 i_2 | i_3\})] = \sum_{r_1 \neq r_2} V^{r_1}_{i_1 i_2} V^{r_2}_{i_3} [3]$$

represents the sum

$$\sum_{r_1 \neq r_2} V^{r_1}_{i_1 i_2} V^{r_2}_{i_3} + \sum_{r_1 \neq r_2} V^{r_1}_{i_1 i_3} V^{r_2}_{i_2} + \sum_{r_1 \neq r_2} V^{r_1}_{i_2 i_3} V^{r_2}_{i_1}.$$

It is natural to group the partitions into classes of the same order, setting $h(\alpha) = \sum_{\Upsilon_\alpha} g(\Upsilon_\alpha)$, where Υ_α represents an arbitrary partition of order α, then

$$V_{i_1 i_2 \ldots i_t} = \sum_{\alpha=1}^{t} \sum_{\Upsilon_\alpha} g(\Upsilon_\alpha) = \sum_{1 \leq \alpha \leq t} h(\alpha).$$

As noted earlier, this representation leads immediately to expressions of order of magnitude. In particular, we have, by a simple extension of preceding arguments that

$$E(V_{i_1 \ldots i_{t_1}} V_{j_1 \ldots j_{t_2}}) = O(n^{min(t_1, t_2)}),$$

a special case of which is $\mathrm{var}(V_{i_1 \ldots i_{t_1}}) = O(n^{t_1})$. This is seen by noting that only classes of equal order in $V_{i_1 \ldots i_{t_1}}$ and $V_{j_1 \ldots j_{t_2}}$ can contribute to the expectation $E(V_{i_1 \ldots i_{t_1}} V_{j_1 \ldots j_{t_2}})$, and the largest contributing class is of order $min(t_1, t_2)$.

Proof of Theorem 6.1

The proof requires the following two assumptions

- Differentiating a'_t does not affect its order of magnitude so that since a_t is $O(n^{1-t})$ then $n^{t-1} a'_t = O(1)$. Note that if the order of magnitude decreases then better rates may be possible.
- For all λ_* in a neighborhood of λ_0, $n^{t-1} \left| \left\{ a'_t(\lambda)(\tau'')^t \right\}' \right| \Big|_{\lambda = \lambda_*} \leq \frac{\sum_{i=1}^n H(x_i)}{n}$, where $H(x_i)$ is a function whose mean exists and is independent of λ_0, and $\tau'' = \partial^2 \tau / \partial \lambda^2$. Recall that τ is defined in (8).

The proof continues from this point by applying magnitude arguments for $U^{(k)}$ from Proposition 6.1, and establishing the magnitude of the remainder terms.

Because λ is univariate the condensed Bhattacharyya basis notation will be used in the proof. Taking the expansion (9) as the starting point our aim is to show that all terms in the Taylor series are $O_p(1)$ and that the remainder is $o_p(1)$. Using the fact (McCullagh, 1987) that

$$(\hat{\lambda}_\psi - \lambda_0) = \left(\frac{V_1}{nE(V_1^j V_1^j)} + O_p(n^{-1}) \right)$$

we may replace $(\hat{\lambda}_\psi - \lambda_0)^k$ by $[V_1/\{nE(V_1^j V_1^j)\}]^k$, since the difference comprises terms of a lower stochastic magnitude.

According to Proposition 6.1, $U_t^{(k)}$ can be re-expressed as $(-a_t' V_t^*)^{(k-1)}$. Since successive differentiations of V_t^* increases its stochastic order of magnitude, but differentiating a_t' leaves its order of magnitude unchanged, it is apparent that the term of highest stochastic order in $U_t^{(k)}$ is $(-a_t' V_t^{*(k-1)})$.

The leading term in $V_t^{*(k-1)}$ for $1 \leq k \leq t$ is

$$(\sum_{r_1 \neq r_2 \neq \cdots \neq r_t} V_1^{r_1} V_1^{r_2} \cdots V_1^{r_t})^{(k-1)}. \tag{A·1}$$

After differentiation, the terms of highest stochastic magnitude are obtained when the maximum number of $V_1^{r_j}$ are differentiated, which produces the term

$$\frac{t!}{(k-1)!} \sum_{r_1 \neq \cdots \neq r_t} V_1^{r_1} \ldots V_1^{r_{(t-k+1)}} V_1^{r_{(t-k+2)}\prime} \ldots V_1^{r_t \prime}. \tag{A·2}$$

Now for this Type 1 model $V_1^{r_1 \prime}$ is the constant τ'', so that (A·2) can be written as

$$\frac{t!}{(k-1)!} (\tau'')^{k-1} \frac{(n-(t-k+1))!}{(n-t)!} \sum_{r_1 \neq \cdots \neq r_{(t-k+1)}} V_1^{r_1} \ldots V_1^{r_{(t-k+1)}}.$$

Note that $(n-(t-k+1))!/(n-t)!$ is $O(n^{(k-1)})$ and that $-a_t'$ is $O(n^{(1-t)})$. Furthermore, from (Serfling, 1980, 5.3.4) it is apparent that

$$(\sum_{r_1 \neq \cdots \neq r_{(t-k+1)}} V_1^{r_1} \ldots V_1^{r_{(t-k+1)}}) = O_p(n^{\frac{1}{2}(t-k+1)}),$$

so that the stochastic order of magnitude of $U_t^{(k)}$ is $O_p(n^{-\frac{1}{2}(t-k-1)})$.

It follows that the first k terms in the Taylor series are equivalent in magnitude to

$$n^{\frac{1}{2}(t-1)}\left(\frac{V_1}{n^{\frac{1}{2}}}\right)^k n^{-\frac{1}{2}k}O_p(n^{-\frac{1}{2}(t-k-1)}) \tag{A·3}$$

By applying the Central Limit Theorem to $V_1/n^{\frac{1}{2}}$ and using the fact that the product of $O_p(1)$ terms is again $O_p(1)$, it follows that $\left(V_1/n^{\frac{1}{2}}\right)^k = O_p(1)$. So to verify that (A·3) has a limiting distribution we need only check that the order of magnitudes add appropriately to give a term that is $O_p(1)$.

For the term in the Taylor series corresponding to $k = t+1$ the same arguments establish the existence of a limiting distribution, except that now the term

$$\left(\sum_{r_1\neq\ldots\neq r_{(t-k+1)}} V_1^{r_1}\ldots V_1^{r_{(t-k+1)}}\right)$$

is vacuous.

The remainder term in (9) is $n^{-\frac{1}{2}(t-1)}(\hat{\lambda}_\psi-\lambda_0)^{t+2}U_t^{(t+2)}(\lambda_*)$ and we must check that this is $o_p(1)$.

Extracting the leading terms from $(\hat{\lambda}_\psi-\lambda_0)^{t+2}$ and $U_t^{(t+2)}$ it follows that the stochastically dominant part of the remainder term is equal in magnitude to

$$n^{\frac{1}{2}(t-1)}\left(\frac{V_1}{n^{\frac{1}{2}}}\right)^{t+2} n^{-\frac{1}{2}(t+2)}n^t\left.\left\{\left(-a_t'\right)\left(\tau''\right)^t\right\}'\right|_{\lambda=\lambda_*}.$$

But this has magnitude

$$O_p(1)n^{-\frac{1}{2}}n^{t-1}\left.\left\{\left(-a_t'\right)\left(\tau''\right)^t\right\}'\right|_{\lambda=\lambda_*},$$

and the second assumption indicates that this entire term is $o_p(1)$ as required.

The essential reason why the remainder term in the Taylor expansion is negligible whereas all the other terms are $O_p(1)$ is that further differentiation above the t^{th} derivative of (A·1) no longer increases its stochastic order of magnitude.

References

Barankin, E. W. (1949). Locally best unbiased estimates. *Ann. Mat. Statist.* **20**, 477–501.

BARNDORFF-NIELSEN, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–22.

BHATTACHARYYA, A. (1946). On some analogues of the amount of information and their uses in statistical estimation. *Sankhyā* **8**, 1–14: 201–218: 315–328.

COX, D. R. & REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc. B* **49**, 1–39.

FISHER, R. A. (1935). The logic of inductive inference. *J. R. Statist. Soc.* **98**, 39–82.

GHOSH, J. K. & SATHE, Y. S. (1987). Convergence of Bhattacharyya bounds - revisited. *Sankhyā, Series A* **49**, 37–42.

GODAMBE, V. P. (1976). Conditional likelihood and unconditional optimal estimating equations. *Biometrika* **63**, 277–284.

LEHMANN, E. L. (1981). An interpretation of completeness and Basu's theorem. *J. Am. Statist. Assoc.* **76**, 335–340.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. New York: Wiley.

LINDSAY, B. G. (1982). Conditional score functions, some optimality results. *Biometrika* **69**, 503–512.

MASE, S. (1977). On the attainment of the supremum of the Bhattacharyya lower bounds for variances of estimators. *Aust. J. Stat.* **19**, 144–51.

MCCULLAGH, P. (1987). *Tensor Methods in Statistics.* London: Chapman and Hall.

MCCULLAGH, P. & TIBSHIRANI, R. (1990). A simple method for the adjustment of profile likelihoods. *J. R. Statist. Soc. B* **52**, 325–344.

MCLEISH, D. L. & SMALL, C. G. (1988). *The Theory and Applications of Statistical Inference Functions* Vol. 44 of *Lecture Notes in Statistics.* New York: Springer-Verlag.

NEYMAN, J. (1959). Optimal asymptotic tests of composite hypotheses. In Granander, U. (Ed.), *Probability and Statistics. The Harold Cramér Volume* pp. 213–234. New York: Wiley.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

SMALL, C. G. & MCLEISH, D. L. (1988). Generalizations of ancillarity, completeness and sufficiency in an inference function space. *Ann. Statist.* **16**, 534–551.

SMALL, C. G. & MCLEISH, D. L. (1989). Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika* **76**, 693–703.

SMALL, C. G. & MCLEISH, D. L. (1991). Geometrical aspects of efficiency criteria for spaces of estimating functions. In Godambe, V. P. (Ed.), *Estimating Functions* pp. 267–276. New York: Oxford University Press.

WATERMAN, R. P. (1993). *Projective Score Methods.* Ph.D. thesis, Pennsylvania State University, University Park, PA.

WATERMAN, R. P. & LINDSAY, B. G. (1994). A simple and accurate method for approximate conditional inference in generalized linear models. *To appear, J. R. Statist. Soc. B.*

Table 1: Comparison of informations in a paired binomial experiment

$m_1 = m_2$	J_1	J_2	$U_c Info$	$\left(\frac{J_1}{U_c Info}\right)^{\frac{1}{2}}$	$\left(\frac{J_2}{U_c Info}\right)^{\frac{1}{2}}$
1	0.123067041	0.119318866	0.119318866	1.015585102	1.000000000
2	0.246134082	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
10	1.230670414	1.228639079	1.228639061	1.000826326	1.000000007
20	2.461340827	2.459359901	2.459359897	1.000402652	1.000000001