Second order information loss due to nuisance parameters: A simple measure

Bruce G. Lindsay
Department of Statistics
Pennsylvania State University

Richard Waterman
Department of Statistics
University of Pennsylvania

7 April 1997

Abstract

This paper is about evaluating the loss of efficiency in maximum likelihood estimation due to the estimation, and subsequent "plugging–in" of nuisance parameters. These effects are necessarily observed at the second order because the information in the efficient score and plug in score are equivalent to first order. We provide a two part partition of the information loss by means of an orthogonal decomposition of the nuisance parameter derivative of the efficient score. One part of the decomposition can be interpreted as the information loss due to the need to bias correct the *m.l.e.*The second part is due to to the increased variability that a plug-in score has compared to the efficient score. We provide several worked examples. Our approximate formula for second order information loss provides correct second order information calculations for the two classic Neyman-Scott problems. We show the relationship of our computational formulas with two types of conditional structures, as well as with statistical curvature in the normal curved exponential family. A final example shows the relevance of this calculation in adaptive estimation.

1 Introduction

It is now nearly 50 years since Neyman and Scott (1948) established that maximum likelihood estimation could have severe failings when estimating a parameter of interest in the presence of many nuisance parameters. In particular, they demonstrated that with infinitely many nuisance parameters, the maximum likelihood estimator could fail to be consistent, and even if consistent, could fail to be the most efficient estimator. We return to this problem to show that with a little mathematical prodding, one can derive simple and meaningful ways of assessing the statistical impact of nuisance parameters.

One can decompose the mean squared error of a maximum likelihood estimator into squared bias and variance. If we do so in a setting in which there is a parameter of interest ψ and a nuisance parameter λ, then the variance and bias for maximum likelihood estimator $\hat{\psi}$ of ψ can be decomposed yet further. Some of the variability and bias would occur even if we knew λ exactly, but there are additional effects due to the fact that λ is unknown, and must be estimated. We will call these latter "nuisance parameter" effects; they are the focus of this paper.

In particular, we wish to show that there is a simple concept of second order information loss in the presence of nuisance parameters. It extends the first order information loss due to λ that can be calculated from the Fisher's information matrix. It is straightforward to calculate and so can be used as a means of assessing the magnitude of the statistical difficulties faced when the nuisance parameter is poorly estimated. Formally, it is created by the evaluation of the information in a set of bias-corrected maximum likelihood equations. It can be interpreted as the asymptotic correlation between the optimal "efficient" score that underlies the first order theory and the actual score equations used for estimation.

We note that our approach focuses on the "invariant" part of the information lost due to nuisance parameters. That is, it is invariant under "interest-preserving" transformations of the parameter space, those that preserve the identity of the parameter of interest. In this sense it reflects information because it does not incorporate those aspects of second order error that are due to the particular parameterization used.

Among other features, we will show that it provides a measure of the "conditionality of the problem". Its relevance is most clear when there are many nuisance parameters in a problem, with only a single parameter of interest, so that the nuisance parameters could severely affect first order inference about the parameter of interest. Thus we will adopt a Neyman and Scott (1948) framework for motivation for some of our focal points. We will also use two of their famous examples to exemplify how our information calculations give the "right" answer.

In our Neyman-Scott type framework, the data will consist of a two way array of independent variables X_{ij} having densities $f(x_{ij};\psi,\lambda_i)$, with rows indexed $i = 1,\ldots,p$, and varying numbers of observations per row, $j = 1,\ldots,n_i$. Thus each row is a "cluster" or "stratum", with a cluster-specific parameter λ_i. In this context, the first order information about ψ in the presence of λ, namely J_1, has the standard form

$$J_1 = n_i I_{11.2} = n_i(I_{11} - I_{12}I_{22}^{-1}I_{21})$$

where I_{ij} is the relevant block of the partitioned Fisher's information matrix. Our objective is demonstrate that there is a meaningful definition of the second order information about ψ in any particular row that has the form:

$$J_2 = J_1 - C_{Bias} - C_{Var} + O(1/n)$$

The terms C_{Bias} and C_{Var} are both $O(1)$ in n_i, and correspond to two distinct costs: C_{Bias} can be interpreted as the cost of bias-correcting the maximum likelihood estimator while C_{Var} arises from the variability of the nuisance parameter estimate. The formulas will be given in Section 2. We will write $J_2^* = J_1 - C_{Bias} - C_{Var}$ for the case when we drop the $O(1/n)$ term.

We note that this information analysis is for the second order information loss of the bias-corrected maximum likelihood estimator. It should be noted that if one models the nuisance parameters λ_i as coming from a distribution, following the proposal of Kiefer and Wolfowitz (1956), then it is possible to reduce the variability cost of the nuisance parameters. See Lindsay (1985) for an extended treatment of how such an empirical Bayes treatment can improve efficiency over the "fixed effects" maximum likelihood estimator we consider here.

Our first task, in Section 2, is to create the asymptotic framework that motivates our work. In Section 3, we consider the bias that arises in maximum likelihood estimation, and develop the natural way to eliminate the bias to the second order, together with the corresponding cost involved with this correction. Sections 4, 5, and 6 consider the implications of the calculations in a variety of models, including those with and without various kinds of conditional structures.

2 Setting up the problem

We start by reviewing some relevant first order asymptotic theory, which will be used to motivate our considerations in the second order theory. We will adopt the convention here of deriving our results in the context of what we will call the "scalar" case, in which both ψ and λ are scalar parameters. This will enable us to avoid complicated arrays, while not significantly changing the nature of the problem. We will write the key formulas in the vector case, however. Many similar calculations have provided in the multidimensional case in Waterman and Lindsay (1996a,b, 1997) which provide much of the foundation of this analysis. A detailed treatment of the calculations as well as background material can be found in Waterman (1993).

2.1 The first order theory

Our initial focus is on the asymptotics in a single row, in which $n_i \to \infty$. In the ensuing developments, when we are doing calculations within a single row, we will drop the strata subscript, i, and write n for n_i and so on. Whenever i does reappear as a subscript, it will refer to the corresponding variable as calculated from row i.

Within a fixed row, define the score function with respect to the parameter of interest, ψ, as $U_0 = \partial_\psi\{\log f(x;\psi,\lambda)\}$ and the score function with respect to λ as $V_1 = \partial_\lambda\{\log f(x;\psi,\lambda)\}$. and denote the operation of projection onto the closed linear space spanned by $\{V_1\}$ by Π_1. That is, $\Pi_1 S$, for any random variable S, equals $\hat{\rho}V_1$, where $\hat{\rho}$ minimizes $E[S-\rho V_1]^2$. The *efficient score* U_1 is obtained by removing from U_0 its projection onto the space spanned by V_1 and therefore has the representation

$$U_1 = U_0 - \Pi_1 U_0.$$

This can be explicitly calculated as

$$U_1 = U_0 - \frac{E(U_0V_1)}{E(V_1V_1)}V_1.$$

Note that $Var(U_1) = J_1$, so that the efficient score reflects the first order information about the parameter ψ in the presence of λ.

This is reflected in the first order asymptotic theory as follows: let $\hat{\lambda}(\psi)$ be the maximum likelihood estimator for λ when ψ is fixed – it can be obtained by solving $V_1(\psi,\lambda) = 0$ in λ. The *plug-in score* is defined to be $\hat{U}_1(\psi) = U_1(\psi,\hat{\lambda}(\psi))$, and the maximum likelihood estimator for ψ can be obtained by equating this score to zero. If we were to estimate ψ and λ from a single stratum, with n going to infinity, then we could analyze the maximum likelihood estimator of ψ through its plug-in score equation $\hat{U}_1(\psi) = 0$. In the first order asymptotic theory for the stratum, we have that the centered maximum likelihood estimator $(\hat{\psi} - \psi)$ is asymptotically equivalent to $J_1^{-1}U_1$. We note that it follows that the maximum likelihood is equivalent to the solution to the efficient score equation $U_1(\psi, \lambda) = 0$, where we have altered the plug-in score equation by replacing the maximum likelihood estimator $\hat{\lambda}(\psi)$ with the true value of λ.

However, the equivalence of inference from the plug-in score and the efficient score fails at the second order. We will later show that $U_1(\psi, \lambda)$ and $\hat{U}_1(\psi)$ are first order equivalent, in that

$$n^{-1/2}[U_1(\psi, \lambda) - \hat{U}_1(\psi)] \rightarrow_{prob} 0$$

In Section 3.2 we look more closely at this convergence and find the key terms that reveal second order effects. In particular, the plug-in score differs from the efficient score in that it is not necessarily mean zero, and it has a variance not necessarily equal to J_1. It is this distinction that causes the second order departure of maximum likelihood estimation from its ideal properties.

In the multistrata model, the same basic observations apply. One can find the maximum likelihood estimator of ψ by solving the plug-in equations:

$$\sum U_{1i}(\psi, \hat{\lambda}_i(\psi)) = 0. \tag{1}$$

If we naively applied first order theory within each row the estimator would be asymptotically equivalent to the solution to the efficient score equations:

$$\sum U_{1i}(\psi, \lambda_i) = 0. \tag{2}$$

The difference between the plug-in score and the efficient score now gets accumulated over the strata, magnifying its effects.

2.2 The formulas

The preceding remarks suggest that a key to the problem of second order information loss is in the dependence of the efficient score on the nuisance parameter. If it depends only weakly on this parameter, then it is intuitively clear that the plug-in score should not deviate much from the efficient score, and so the cost should be low.

This notion can be made more precise by defining the random vector $W = U_1'$, where $'$ refers to differentiation with respect to the nuisance parameter. It can be verified that W has mean zero—it follows because U_1 is orthogonal to V_1. We will call W the *cost score.* This variable clearly determines the nature and the extent of the local dependence of the efficient score on λ, and in particular $W = 0$ if the efficient score does not depend on the nuisance parameter.

If the parameters ψ and λ are both vectors, the cost score is the $\dim(\psi) \times \dim(\lambda)$ matrix $W = \nabla_\lambda U_1^T$.

We can decompose W by projection into two uncorrelated terms $W_b = \Pi_1 W$ and $W_v = W - W_b$. In the scalar case, we let $\kappa = E[WV_1]/E[V_1^2]$, so that

$$W_v = W - \kappa V_1 \text{ and } W_b = \kappa V_1.$$

We define the *bias cost* to be:

$$C_{Bias} = \frac{1}{2} E(W_b I_{22}^{-1} W_b^T),$$

and note that it is largest when the cost score is highly correlated with the nuisance score V_1. In the scalar case, we can reexpress this as $C_{bias} = \kappa^2/2$. This will be the cost of bias-correcting the maximum likelihood estimator.

On the other hand, the *variance cost* will be

$$C_{Var} = E(W_v I_{22}^{-1} {W_v}^T),$$

which is large if W has a large component orthogonal to the nuisance score space. This is the cost in estimation due to the additional variability in the plug in score over of the efficient score.

We note the curiosity that the overall information cost $C_{bias} + C_{var}$ is off by the factor of $1/2$ in C_{bias} from simply being the normalized variance of the cost score W. Our intuition is that some of the bias cost of the *m.l.e.* gets recovered by using bias correction.

The simplicity of these formulae makes them a handy tool for a preliminary analysis of the secondary costs in models with many parameters, including semiparametric models. In addition to measuring the overall cost, the magnitude of the separate components separately indicate potential problems with bias and inflated variance. We do note, however, that these simple formula arise from disregarding terms of smaller order, and so can give negative informations when calculated with small values of n. We provide an extensive set of examples later in the paper; we first provide some further theoretical justification for their relevance.

2.3 Information in estimating functions.

The starting point of our strategy is that the maximum likelihood equation for ψ, as given in equation (1), has the form $H(\psi) = 0$ for a particular function H. This links our subject to that of estimating functions, the theory of which we now briefly review.

We will start by considering information concepts in the simplest case, that is when there are no nuisance parameters. For an estimating function, say $H(\psi)$ that is mean zero, we define the information for vector ψ as

$$\mathrm{Info}(H) = E(\nabla_\psi H) E(HH^T)^{-1} E(\nabla_\psi H)^T. \tag{3}$$

If the function H is sum of independent terms, then the usual theory of M-estimation indicates that the asymptotic variance of the estimator satisfying

$H(\psi) = 0$ is equal to the inverse of the information, which gives it statistical content as being a measure of information.

Following the arguments of Godambe (1960), differentiation of the mean zero identity $E_\psi[H(\psi)] = 0$ with respect to ψ leads to the information identity $E(\nabla_\psi H + HU_0) = 0$. If we then replace $E(\nabla_\psi H)$ by $-E(HU_0)$ in (3), we can recognize that in the scalar case, the information equals the Fisher information $E(U_0^2)$ times the squared correlation of H and U_0. It immediately follows that the information in an estimating equation H is maximized when it is a linear function of U_0.

On the other hand, if there exists a nuisance parameter λ, and we are evaluating a zero-unbiased function $H(\psi)$ for its information, we can differentiate $E_{\psi,\lambda}[H(\psi)]$ with respect to λ and obtain $E[\nabla_\lambda H + HV_1] = 0$. Since H is free of λ, the last equality shows that H is orthogonal to V_1, which in turn implies that we can measure information by correlation with the efficient score:

$$\text{Info}(H) = E(U_1 H^T) E(HH^T)^{-1} E(HU_1^T).$$

This is the formula we will later use to calculate information.

2.4 The strategy

We return to the estimating equation for the maximum likelihood estimator:

$$\sum U_{1i}(\psi, \hat{\lambda}_i(\psi)) = 0.$$

If the plug-in scores $\hat{U}_{1i}$ were mean zero, then one could apply the foregoing theory of estimating functions (or M-estimation) to derive the asymptotic distribution of the solution as $p \to \infty$. However, there is too much bias in these equations to use them directly.

The first part of the strategy is to replace the plug-in score $\hat{U}_1(\psi)$ with a bias-corrected version $\hat{U}_2(\psi) = U_2(\psi, \hat{\lambda}_\psi)$. When this is done, solving

$$\sum \hat{U}_{2i}(\psi) = 0$$

gives an estimating function that produces a bias-corrected maximum likelihood estimator, and the bias in the scores is sufficiently reduced that one can proceed with an asymptotic information calculation, as described above. Thus, we desire to calculate Info$(\hat{U}_2)$.

Since the presence of the maximum likelihood estimator $\hat{\lambda}_\psi$ in $\hat{U}_2$ makes calculating Info$(\hat{U}_2)$ potentially tedious, and because we wish simple formulas for the second order in n, we create a surrogate estimating function $U_2^\dagger$ that is information equivalent, in the sense that Info$(\hat{U}_2)$ =Info$(U_2^\dagger) + O(1/n)$. Finally, we simplify yet further by showing that Info$(U_2^\dagger) = J_2^* + O(1/n)$.

Remark. An alternative to this strategy would be to analyze the asymptotic mean square of the maximum likelihood estimator in some appropriate fashion that separates out the invariant aspects. We have chosen our present course because we believe it is more straightforward. However, we could here follow the decomposition of the mean squared error of the bias-corrected maximum likelihood estimator such as developed by Efron (1975) and later advanced by Amari (1982).

If we did so in the asymptotic setting where there are a finite number of strata and the stratum sample sizes becoming infinite at equal rates, the resulting mean squared error for ψ would have components due to parameterization effects and statistical curvature. With a slight modification, the inverse of the information we calculate here can be shown equal (to the right order in n) to the "invariant" components of mean squared error – those that do not depend on the choice of the ψ parameterization. The distinction is that C_{var} must be augmented with a correction term that is $O(1/p)$, where p is the number of strata, and so washes out in an infinite strata calculation. However, it could be significant for a small number of strata, especially as to the case of a single stratum. Of our two components, C_{bias} corresponds to parameterization effects that are unavoidable because the parameter of interest must be preserved and C^*_{var} corresponds to the effects of statistical curvature.

The second order optimality of the bias corrected *m.l.e.* then establishes that the information quantities we calculate here, as adjusted above, provide second order lower bounds for information loss.

3 Correcting the *m.l.e.* for bias

We start with a familiar example from illustrate the key issues in bias correction.

3.1 The many means example.

We suppose that the i^{th} stratum consists of n_i independent observations from the normal density with parameters (μ_i, σ^2). When viewed across the strata, we have many means μ_i and a single variance σ^2. The σ^2 parameter will be considered the parameter of interest, and so corresponds to ψ.

Before we turn to our bias arguments, we illustrate some of our basic calculations on this example. It is straight forward to calculate the scores in the i^{th} stratum as

$$\begin{aligned}\frac{\partial \log f(x_i;\sigma^2,\mu_i)}{\partial \sigma^2} = U_0 &= -\frac{n}{2\,\sigma^2} + \frac{1}{2\,\sigma^4}\sum_{j=1}^{n}(x_{ij}-\mu_i)^2 \\ \frac{\partial \log f(x_i;\sigma^2,\mu_i)}{\partial \mu_i} = V_1 &= \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_{ij}-\mu_i).\end{aligned}$$

We can therefore calculate the information matrix to be:

$$I = \begin{bmatrix} \frac{n}{2\sigma^4} & 0 \\ 0 & \frac{n}{\sigma^2} \end{bmatrix}.$$

Direct calculation then shows that $U_1 = U_0$, and so

$$W = U_1' = -\frac{1}{\sigma^4}\sum_{j=1}^{n}(x_{ij} - \mu_i).$$

We can then calculate $\kappa = E[V_1 W]/E[V_1^2] = -1/\sigma^2$. Because W is perfectly correlated with V_1, we must have $W_v = 0$ and $C_{var} = 0$. We also have $C_{bias} = \kappa^2/2 = \frac{1}{2\sigma^4}$. It follows that

$$J_2^* = \frac{n}{2\sigma^4} - \frac{1}{2\sigma^4} = \frac{n-1}{2\sigma^4}.$$

Thus our approximate information calculation correctly recovers the degree of freedom lost by estimating the mean.

We return to the bias problem. If the expectation of U_0 is taken when μ_i is replaced by it maximum likelihood estimator, $\overline{x}_i$, then $E\{U_0(\sigma^2, \overline{x}_i)\} = -1/(2\sigma^2)$ indicating that the plug-in score is biased. The score bias accumulates over the strata to such an extent that the maximum likelihood estimator for σ^2 is inconsistent when the stratum sample sizes are held fixed while $p \to \infty$. In the case where all the n_i are equal to n, $\hat{\sigma}^2 \to \{(n-1)/n\}\sigma^2$. Thus the asymptotic bias is substantial for small n.

Waterman and Lindsay (1997) proposed a *rectangular array asymptotics* to get a closer understanding of the relative roles of the magnitudes of n_i and p. To illustrate their approach, if one used an asymptotics in which there are p rows, each with γp observations, then as $p \to \infty$, the maximum likelihood estimator is still consistent, but it has the limiting property that $\sqrt{N}(\hat{\sigma}^2 - \sigma^2)$, where N is the total sample size $\gamma p \cdot p$, is normal with mean $-\gamma^{-1}\sigma^2$ and variance $2\sigma^4$. That is, bias shows up in the mean of the limiting distribution. However, the asymptotic variance is exactly as it would be for a regular problem.

This example raises the question as to whether there is a systematic way to eliminate the bias at this level of asymptotics. We next review the methodology developed by Waterman and Lindsay (1996a) that achieves this bias elimination within the context of estimating functions.

3.2 Bias Removal and the U_2 score

The efficient score U_1 generally has magnitude $O_p(n^{1/2})$. Although U_1 is mean zero, the substitution of the nuisance parameter estimate generally induces a bias of order $O(1)$ in the plug-in score. This bias is such that it dominates in the calculation of mean squared error unless its magnitude is reduced to $O(n^{-1/2})$. Fortunately there is a natural extension of the preceding developments that leads to an estimating equation with the appropriate degree of bias removal.

We define the second order Bhattacharyya scores, for vector λ case, to be:

$$V_{2(ij)} = \frac{\frac{\partial^2}{\partial\lambda_i\partial\lambda_j} f(x;\psi,\lambda)}{f(x;\psi,\lambda)}.$$

Higher order scores, such as $V_{3(ijk)}$, can be similarly defined, but are not needed here. Returning to the scalar λ case, we define the projection operator Π_2 to be the action of projection onto the space spanned by V_1 and V_2. The second order projected score is defined to be $U_2 = U_0 - \Pi_2 U_0$.

It was shown by Small and McLeish (1989) that the second order score eliminates bias in the sense that $U_2(\psi, \hat{\lambda}_\psi) = \hat{U}_2$ has bias of order $O(n^{-1/2})$. We offer a summary of the necessary calculations to establish the bias results, as well as identify other second order effects.

First, since both nuisance score V_1 and the variance component of the cost score, W_v are i.i.d. sums, and they are orthogonal, one can determine the joint limiting distribution to be:

$$nm_{11}^{-1/2}\begin{pmatrix} W_v \\ V_1 \end{pmatrix} = \begin{pmatrix} Z_{1n} \\ Z_{2n} \end{pmatrix} \to_{dist} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},$$

where Z_1 and Z_2 are independent mean zero normal variables with respective variances C_{var} and 1.

With the help of these variables, we can describe more precisely how the plug-in score $\hat{U}_1$ departs from the efficient score $U_1$by capturing their difference to the next order of magnitude. It can be shown that

$$\hat{U}_1 - U_1 - Z_{1n} Z_{2n} + \frac{\kappa}{2} Z_{2n}^2 \rightarrow_{prob} 0.$$

The proof is a straightforward asymptotic expansion argument. Here the term κ is as defined earlier, equal to $E[WV_1]/E[V_1^2]$, with $C_{bias} = \kappa^2/2$. We can rewrite this convergence in the form:

$$\hat{U}_1 - U_1 \rightarrow_{dist} Z_1 Z_2 + \frac{\kappa}{2} Z_2^2$$

Thus we see that if κ, and hence C_{bias}, is 0, then the asymptotic mean of $\hat{U}_1 - U_1$ is zero, but otherwise there is a bias term of magnitude $\kappa/2$. Notice that this second order convergence implies in an elementary fashion the first order equivalence

$$n^{-1/2} \left[\hat{U}_1 - U_1 \right] \rightarrow_{prob} 0.$$

The score function $\hat{U}_2$ very precisely removes the bias term in the sense that

$$\hat{U}_2 - U_1 \rightarrow Z_1 Z_2 + \frac{\kappa}{2} Z_2^2 - \frac{\kappa}{2},$$

and so gives a score function that is closer on average to the true efficient score U_1. However, if we want to represent $\hat{U}_2$ with a simpler score type variable, we could do so with less variation by using U_2, because:

$$\hat{U}_2 - U_2 \rightarrow Z_1 Z_2.$$

If there is no variance cost, so that $C_{var} = 0$, we must have Z_1 degenerate at zero, and so in this case the $\hat{U}_2$ score is asymptotically equivalent to U_2 at the second order, and we could use $\mathrm{Info}(U_2)$ for second order information, as we establish in the next section.

However, in general C_{var} is not zero, in which case U_2 is not sufficiently close to $\hat{U}_2$ to accurately reflect information to the second order, so we first find a better surrogate.

3.3 The surrogate estimating function

The above asymptotic statements lead us directly to a method for constructing a simple surrogate for the estimating function $\hat{U}_2$, as we have

$$\hat{U}_2 - \{U_1 + Z_{1n}Z_{2n} + \frac{\kappa}{2}[Z_{2n}^2 - 1]\} \rightarrow_{prob} 0.$$

Thus we identify the term in brackets, written in terms of scores,

$$U_2^\dagger = U_1 + \frac{V_1 W_v}{n\, I_{22}} + \frac{\kappa}{2}\left[\frac{V_1^2}{nI_{22}} - 1\right].$$

as a simpler surrogate for $\hat{U}_2$.

The next proposition shows that if an estimating equation, H, is close to the efficient score U_1, then the information in H has a simple second order form: (The subscript n is introduced to emphasize the fact that the functions under consideration depend on n and that the limits are taken as n approaches infinity.)

Proposition **3.1** *Assume that the estimating function H_n can be written as*

$$H_n = U_{1n} + \Delta_n,$$

and that the following two conditions are satisfied: that

$$E[\Delta_n^2] \rightarrow C$$

and that

$$\frac{E[U_{1n}\Delta_n]}{E[U_{1n}^2]} \rightarrow 0.$$

Then Info$(U_{1n})-$ *Info*$(H_n) \rightarrow C$.

Proof: by the definition of information and from the construction of H_n we have that

$$\mathrm{Info}(H_n) = \frac{E^2(U_{1n}H_n)}{E(H_n^2)} = \frac{E^2\{U_{1n}(U_{1n}+\Delta_n)\}}{E\{(U_{1n}+\Delta_n)^2\}},$$

which can be rewritten as

$$\frac{[E\{(U_{1n}+\Delta_n)^2 - \Delta_n(U_{1n}+\Delta_n)\}]^2}{E\{(U_{1n}+\Delta_n)^2\}}.$$

Expanding the square in the numerator and canceling by the denominator gives

$$E\{(U_{1n}+\Delta_n)^2\} - 2E\{\Delta_n(U_{1n}+\Delta_n)\} + \frac{E^2(\Delta_n U_{1n}+\Delta_n^2)}{E\{(U_{1n}+\Delta_n)^2\}}.$$

The third term is $o(1)$ since the numerator is at most $O(1)$ whilst the denominator is $O(n)$. On simplifying the first two terms we obtain

$$\mathrm{Info}(H_n) = E(U_{1n}^2) - E(\Delta_n^2) + o(1),$$

which is the desired result. ■

We can apply this proposition as follows. In our case, we can write

$$\Delta_n = \hat{U}_2 - U_1 = Z_{1n}Z_{2n} + \frac{\kappa}{2}[Z_{2n}^2 - 1] + o_p(1)$$

and so under some further regularity, we have

$$E(\Delta_n^2) \to E\left[Z_1 Z_2 + \frac{\kappa}{2}[Z_2^2 - 1]\right]^2 = C_{bias} + C_{var}.$$

This leads us to the desired conclusion: $\mathrm{Info}(\hat{U}_2) = J_2^* + o(1)$.

4 Type I conditional problems

We next consider the interpretation of our calculations in an important class of examples for which there is an alternative method for bias reduction.

In a subset of nuisance parameter problems there exists a complete and sufficient statistic, say S, for the nuisance parameter λ_i when ψ is fixed, where S does not depend on ψ. We call such a problem a Type I conditional problem to distinguish it from the case when there exists a complete and sufficient statistic $S(\psi)$ that does depend on ψ, which we call Type II. The *many means* problem has Type I conditional structure, with the sufficient statistic for the mean μ_i being $S_i = \overline{x}_i$, which does not depend on $\psi = \sigma^2$.

In all Type I models one can create a conditional score function $U_c(\psi)$ by differentiating the logarithm of the conditional density of the data given S. The conditional score U_c is free of the nuisance parameters and is mean zero and so provides an unbiased estimating function, the solution to which is called the conditional maximum likelihood estimator. The conditional score can be written as

$$U_c = U_0 - E[U_0|S].$$

The conditional maximum likelihood was developed extensively by Andersen (1970). In the *many means* problem, the corresponding conditional maximum likelihood estimator is exactly the bias corrected maximum likelihood estimator $n/(n-1)\hat{\sigma}^2$ when the stratum sample sizes are equal to n.

In such problems it seems reasonable to presume that the conditional score is optimal, and so our J_2 formula should agree with the conditional information to the right order of magnitude. (For more on the optimality of conditional methods, see Lindsay (1983), in which it is established that even if the nuisance parameters are assumed to come from a distribution, it is not possible to construct more efficient estimators.)

Waterman and Lindsay (1996a,b) extensively considered the relationship of the corrected score U_2 and the conditional score U_c. We briefly summarize some of their key findings.

The first surprise is that in every Type I model with exponential family structure, there is in fact no variance cost associated with using $\hat{U}_2$; that is, $W = U_1'$ is a linear function of V_1 so that $U_1' - \kappa V_1 = 0$. (This is a special case of a more general result that says that the tth derivative of U_t is a linear function of V_t^*, where V_t^* is an orthogonalized version of the tth Bhattacharyya basis function.)

It follows that $\hat{U}_2$ is asymptotically equivalent to U_2. The information in U_2 was calculated by Waterman and Lindsay and shown to satisfy:

$$\text{Info}(U_2) = J_1 - C_{bias} + O(1/n) = J_2^* + O(1/n).$$

They also establish the relationship between the U_2 score and the conditional score, showing that $E(U_c - U_2)^2$ is $O(1/n)$, so that Info(U_c) =Info(U_2) + $O(1/n)$. Thus J_2^* captures the information in U_c to second order.

This establishes our basic information formula as being appropriate for capturing the second order information in Type I conditional problems. As a simple illustration, note that in the *many means* example, the information in the conditional score function equals $J_1 - C_{Bias} = \frac{n-1}{2\sigma^4}$ exactly, capturing the loss of a degree of freedom from estimating the nuisance parameter. We conclude this section with some further illustrations of the calculations.

4.1 Paired Poisson.

In some special cases the U_1 score function is equal to U_c. An example of this is the paired Poisson problem, where we observe two independent Poissons with means $\exp(\psi + \lambda)$ and $\exp(\lambda)$ respectively. Here, $X + Y$ is the complete sufficient statistic for λ and $V_1 = x + y - E(X+Y)$. The conditional distribution of X given $X + Y$ is well known to be Binomial with parameters $N = X + Y$ and $p = \exp(\psi)/1 + \exp(\psi)$. Using this fact it is easy to show that

$$E(U_0 \,|\, X + Y) = (X + Y)\frac{\exp(\psi)}{1 + \exp(\psi)} - \exp(\psi + \lambda).$$

This is a linear function of V_1, so that $U_1 = U_c$ and therefore $J_1 = J_2^*$ =Info(U_2), so there is no second order cost to estimation of the nuisance parameter.

4.2 Paired Bernoulli

Moving now to the building block for many commonly used statistical models (e.g., logistic regression, Rasch models), we consider a paired Bernoulli experiment. We observe two independent Bernoulli variables with odds of success $\exp(\psi + \lambda)$ and $\exp(\lambda)$ respectively, so that the parameter of interest is the log-odds ratio ψ. This is again a Type I model with $X + Y$ being the complete and sufficient statistic for λ. One interesting aspect of this model is that $U_2 = U_c$.

It can be shown, preferably by symbolic computation, that

$$\kappa = \frac{2\, e^{\lambda+\psi} \left(1 + e^{\lambda}\right) \left(e^{\psi} - 1\right) \left(1 + e^{\lambda+\psi}\right)}{\left(1 + e^{\psi} + 4\, e^{\lambda+\psi} + e^{2\,(\lambda+\psi)} + e^{2\,\lambda+\psi}\right)^2}.$$

Clearly κ is equal to zero when $\psi = 0$, and in this situation U_1 also equals $U_2 = U_c$. One can indeed verify our earlier claim that $U_1' = \kappa V_1$, so $W_v = 0$, although the details of this verification are rather complicated. So, as predicted, the only cost is $C_{bias} = \frac{1}{2}\kappa^2$.

To get a feeling for the speed in which the second order asymptotics kicks in we can let the Bernoulli grow into a Binomial. The following calculations are for paired Binomial experiments with sample sizes of 1 through 5, 10, and 20. All calculations are exact, and the parameters are set as $\psi = 0.5$ and $\lambda = -0.25$.

INSERT TABLE 1 HERE

We note that our second order approximation only captures half the information loss for $n = 1$, but improves rapidly and quite generally captures the overall magnitude.

In this particular example it could be argued that the second order information loss is negligible and that from a practical perspective of no consequence. However, the information loss is a function of the curvature κ, which for Bernoulli problems is an increasing function of the asymmetry of the probabilities. If we had considered parameter values for $\psi = 2$ and $\lambda = -1$, then the previous table would have had a first line of

INSERT TABLE 2 HERE

so that the second order loss in information would have been of the order of 50%.

5 Type II conditional models

We turn now to conditional models that display a different kind of structure because the sufficient statistic $S(\psi)$ for the nuisance parameter depends on the parameter of interest. There is an interesting subclass of these models which are in the exponential family and characterized by densities of the form

$$f(x, \psi, \lambda) = \exp\{\lambda S(\psi) - \lambda\tau(\psi, \lambda)\}.$$

Such a structure arises when the parameter of interest ψ is the ratio of natural parameters. In this case, the conditional score function depends on the nuisance parameter, but only through a weighting factor ($'$ denotes differentiation with respect to ψ):

$$U_c = \lambda[S'(\psi) - E_\psi\{S'(\psi) \mid S(\psi)\}].$$

Notice the important point that with this type of conditioning, the conditional score depends on λ, and we will find that in this class, unlike the Type I, we will have $C_{var} > 0$.

One can show that in this class of models, U_1 is equal to U_c, and that W, its derivative with respect to lambda is therefore free of lambda. Furthermore, W is orthogonal to V_1 so that $C_{Bias} = 0$ and the information loss is therefore entirely in C_{var}. We consider some important examples with this kind of Type II structure.

5.1 The many variances problem

We return to a second important Neyman-Scott example. Suppose now that each stratum consists of n_i independent normal variables with common mean μ and variance σ_i^2 that depends on the stratum. We now consider the mean parameter to be of interest, and the nuisance parameters are the many variances σ_i^2.

We have earlier laid out in the *many means* problem the necessary first order calculations—one only needs to reverse the role of the two parameters. The U_0 score function is therefore

$$U_0 = \frac{1}{\sigma_i^2} \sum_{j=1}^{n_i} (x_{ij} - \mu).$$

It is straightforward to show that because U_0 is linear in μ and the nuisance parameter enters only as a weight that $U_0 = U_1 = U_2$. Furthermore, one can show $U_0 = U_c$. We can further calculate that

$$W = -\frac{1}{\sigma_i^4} \sum_{j=1}^{n_i} (x_{ij} - \mu).$$

Since W is orthogonal to V_1, the score for σ^2, it follows that $C_{bias} = 0$, and we can calculate $C_{var} = E[W^2]/I_{22} = 2/\sigma^2$. Therefore the corrected information about μ is

$$J_2^* = \frac{n}{\sigma^2} - \frac{2}{\sigma^2} = \frac{n-2}{\sigma^2}.$$

We can compare this with the estimating function analysis of the actual maximum likelihood estimator. In this case, the plug-in score is unbiased, $E\{U_0(\mu, \hat{\sigma}^2_{i\,\hat{\mu}})\} = 0$, so that one can proceed to show that the maximum likelihood estimator of μ is consistent. One can therefore apply estimating function methodology to find its information, which for a single stratum is $\text{Info}(\hat{U}_0) = J_2^*$, so our second order formula is exact in this case.

5.2 Paired exponentials.

For another example from this class of models, consider two exponential random variables with means $E(X) = \frac{1}{\psi\lambda}$ and $E(Y) = \frac{1}{\lambda}$. A complete sufficient statistic for the nuisance parameter λ is given by $\psi X + Y$, which depends on ψ; our characterization of a Type II problem. The score functions U_0 and V_1 are

$$U_0 = \frac{1}{\psi} - \lambda X,$$

$$V_1 = \frac{2}{\lambda} - (\psi X + Y).$$

The U_1 score function is simply

$$\lambda \left(\frac{y - \psi x}{2\psi} \right),$$

so that W is simply the term within the brackets. Once again $W = W_v$ and a straight forward calculation shows that

$$E(W_v^2) = \frac{1}{2\psi^2\lambda^2},$$

so that C_{Var} is equal to $1/4\psi^2$. The first order information $J_1 = n/2\psi^2$, so that

$$J_2^* = \frac{n - \frac{1}{2}}{\psi^2},$$

reflecting the loss in some sense of "half a degree of freedom".

In this model there is a marginal distribution that we can exploit to obtain an exact calculation of the information in the maximum likelihood estimator, which is based on $\hat{U}_1$. The key fact is that the marginal likelihood of $X./Y.$, where . denotes summation, is free of λ. The information in $\hat{U}_1$ can be calculated as $-E\{(\partial/\partial\psi)\hat{U}_1\}$ because this is the score function in the marginal likelihood and therefore information unbiased. One can show that

$$\frac{\partial}{\partial\psi}\hat{U}_1(\psi) = -\frac{2n}{\psi^2}\frac{\psi X.}{(\psi X. + Y.)^2},$$

and noting that the random variable $\frac{\psi X.}{(\psi X.+Y.)^2}$ has a Beta distribution with parameters (n,n) allows the information to be readily calculated as

$$E\left(-\frac{\partial}{\partial\psi}\hat{U}_1(\psi)\right) = \frac{n-\frac{n}{2n+1}}{2\psi^2}.$$

Thus we see that once again J_2^* has captured the information terms to the right order, although it is not exact in this instance.

6 Models without conditional structure

We next turn to models that lack conditional structure, but by their simple nature can offer us a new set of insights into the loss from nuisance parameters. We first consider models for pairs of independent normal variables with common known variance. The nuisance parameter effects arise because parameters ψ and λ are nonlinear functions of the normal means.

6.1 Circular normal

We illustrate with a simple example. The stratum consists of a pair of random variables, X and Y, with distributions $N(\rho\cos(\phi),1)$ and $N(\rho\sin(\phi),1)$. We alternatively consider the parameter of interest to be radius ρ and angle ϕ. (The model corresponds to a bivariate normal distribution with the unknown mean being parameterized in polar coordinates.)

The two score functions are $U_\rho = r\cos(\hat{\phi}-\phi)-\rho$ and $U_\phi = \rho r\sin(\hat{\phi}-\phi)$, where r and $\hat{\phi}$ are the observed values of the radius and angle respectively. The Fisher's information matrix is

$$\begin{pmatrix} I_{\rho\rho} & I_{\rho\phi} \\ I_{\phi\rho} & I_{\phi\phi} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \rho^2 \end{pmatrix}.$$

The parameters are orthogonal, (the off diagonal elements of the Fisher's information matrix are zero) which makes the calculations simpler, in particular $U_1 = U_0$ for both versions of the parameter of interest.

First consider the case where ρ is the parameter of interest, ψ. The Fisher's information for n observations is $J_1 = n \cdot 1$, and $W = -r sin(\hat{\phi} - \phi)$, which is perfectly correlated with V_1, so $C_{Bias} = 1/(2\rho^2)$ and $C_{Var} = 0$. In particular, we have

$$J_2^* = n - \frac{1}{2\rho^2}.$$

Thus we will have bias problems with maximum likelihood in this setting, and they will be aggravated as ρ becomes small, but there is no further variance cost in estimating the parameter ρ.

On the other hand, if ϕ is the parameter of interest, we can calculate $J_1 = \rho^2$ and $W = r\sin(\hat{\phi} - \phi)$. The latter is orthogonal to V_1 so $C_{Bias} = 0$ and $C_{Var} = 1$. This gives

$$J_2^* = n\rho^2 - 1.$$

and indicates that a multistrata *m.l.e.* will not be bothered by bias, but will have severe variability problems if the nuisance parameters ρ_i are sufficiently small.

An alternative way of comparing the information calculations is to consider the *relative information*, defined here as J_2^*/J_1. For the two versions of the circular normal problem the relative information is $1 - 1/(2\, n\, \rho^2)$ and $1 - 1/(n\, \rho^2)$ respectively, which clearly shows the role of the curvature.

6.2 Adaptive estimation

A curious feature of the location problem is the phenomenon of adaptive estimation. It has been shown by Bickel (1982) that if we have the semiparametric model $f(x - \theta)$ where f is an unknown density that is symmetric about zero, then it is possible to estimate θ as accurately with f unknown as with f known. That is, it is possible to achieve the asymptotic variance formula $E[f'(x)/f(x)]^2$, here and hereafter setting $\theta = 0$ to simplify formulas.

We can here think of this as a nuisance parameter problem, with the parameter of interest being θ, and the nuisance parameter being the infinite dimensional space of functions $f(\cdot)$. The concept of efficient score has been adapted to such a semiparametric setting, and in this case the efficient score is $U_1 = -f'(x)/f(x)$, which also equals U_0 (the prime here refers to x differentiation). Thus, based on first order information calculations, there is no cost to not knowing the function-valued parameter f.

However, it is clear that U_1 depends very strongly on the nuisance function, so we would expect some second order costs. One simple way to consider the higher cost involved from the estimation of the density f is to consider a simple nuisance parameter model of the form

$$h(x, \theta, \pi) = (1 - \pi) f(x - \theta) + \pi g(x - \theta),$$

where f and g are two symmetric densities, and the mixing parameter π is the nuisance parameter. The efficient score is here

$$U_0 = U_1 = \frac{-h'}{h}$$

The nuisance parameter score is

$$V_1 = \frac{g - f}{h}.$$

The cost score is

$$W = -\partial_\pi \frac{h'}{h} = \frac{h \cdot (g' - f') - h'(g - f)}{h^2} = \frac{fg' - f'g}{h^2}.$$

In this case, $C_{Bias} = 0$, but $C_{Var} = E_h[(U_f - U_g)gf/h^2]^2/E[V_1^2]$.

As an illustration of these calculations, if f is standard normal and g is standard Cauchy, which are rather dramatically different, then at $\pi = \frac{1}{2}$, we can calculate numerically that $J_1 = n(.67)$, whereas $C_{var} = .23/.27 = .85$, corresponding to only the loss of about $.85/.67 = 1.27$ observations at the second order. One might compare this with our earlier calculation in the many variances problem that showed a second order cost of 2 observations as the price for not knowing the normal variance.

7 Discussion

We have shown that by decomposing the variance of the cost score W, one can obtain simple and meaningful measures of the second order costs of nuisance parameters. This leaves open the question of whether they could fruitfully be used in a more direct manner, such as a route to improving confidence interval statements. We rather suspect that is it so, but have our doubts that in the modern age, such second order improvements would compete well against more computer-intensive methods, such as a parametric bootstrap, which would incorporate all the aspects of finite-sample properties, not just the ones we have focussed on.

We rather think that our analyses are more fruitfully used for understanding several more theoretical questions.

One, which we raised in our last example, relates to whether these methods could be fruitfully used to gain further understanding of the costs of joint estimation in semiparametric models, where the dimension of the nuisance parameter could be sufficiently large as to raise very large second order costs.

Another question relates to modeling a large set of nuisance parameters as coming from a random effects distribution as opposed to a fixed effects Neyman-Scott type structure: when is this beneficial? This is a case where the benefits in a single stratum do not come at the first order, but perhaps a second order analysis incorporating this aspect could provide some insights.

In our analysis we have used the natural (from an analytic viewpoint) estimator for λ. By definition we are not interested in λ, so would be happy if we could use a simpler estimate for it; two immediate candidates are method of moments estimates and one step approximations to the maximum likelihood estimator. If we are in a random effects environment, we could used posterior expectations of the nuisance parameters. We are interested in the efficiency consequences of such methods.

Another area of interest is the subject of approximate conditionality. We know that in Type I models, those in which conditional distributions exist that are free of the nuisance parameters, then $C_{Var} = 0$. We are interested to explore the extent to which the statement $C_{var} = 0$, or equivalently $W \perp V_1$, is a substitute for conditionality.

References

[1] S Amari. Differential geometry of curved exponential families – curvatures and information loss. Ann. Statist., 10:357–385, 1982.

[2] EB Andersen. Asymptotic properties of conditional maximum likelihood estimators. J. R. Statist. Soc. B, 32:283–301, 1970.

[3] PJ Bickel. On adaptive estimation. Annals of Statistics, 10:647–671, 1982.

[4] B Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). Ann. Statist., 3:1189–1117, 1975.

[5] VP Godambe. An optimum property of regular maximum likelihood estimation. Ann. Mat. Statist., 31:1208–1211, 1960.

[6] J Kiefer, J Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. Ann. Mat. Statist., 27:887–906, 1956.

[7] BG Lindsay. Efficiency of the conditional score in a mixture setting. Ann. Statist., 11:486–497, 1983.

[8] BG Lindsay. Using empirical partially Bayes inference for increased efficiency. Ann. Statist., 13:914–931, 1985.

[9] J Neyman, EL Scott. Consistent estimates based on partially consistent observations. Econometrica, 16:1–32, 1948.

[10] CG Small, DL McLeish. Projection as a method for increasing sensitivity and eliminating nuisance parameters. Biometrika, 76:693–703, 1989.

[11] RP Waterman. Projective Score Methods. PhD dissertation, Pennsylvania State University, University Park, PA, 1993.

[12] RP Waterman, BG Lindsay. The accuracy of projected score methods in approximating conditional scores. Biometrika, 83:1,14, 1996.

[13] RP Waterman, BG Lindsay. A simple and accurate method for approximate conditional inference in generalized linear models. J. R. Statist. Soc. B, 58:177,188, 1996.

[14] RP Waterman, BG Lindsay. Projected score methods for nuisance parameters: Asymptotics and Neyman-Scott problems. Unpublished manuscript, 1997.

Table 1: Information calculations for a paired Binomial experiment. Parameter values are $\psi = 0.5$ and $\lambda = -0.25$.

Trials	J_1	J_2	$J_1 - J_2$	$C_{bias} = \frac{1}{2}\kappa^2$
1	0.123067	0.119319	0.00374818	0.00193296
2	0.246134	0.243584	0.00255057	0.00193296
3	0.369201	0.366896	0.00230507	0.00193296
4	0.492268	0.490069	0.00219923	0.00193296
5	0.615335	0.613195	0.00214026	0.00193296
10	1.23067	1.22864	0.00203133	0.00193296
20	2.46134	2.45936	0.00198093	0.00193296

Table 2: Information calculations for a paired Bernoulli experiment. Parameter values are $\psi = 2.0$ and $\lambda = -1.0$.

Trials	J_1	J_2	$J_1 - J_2$	$\frac{1}{2}\kappa^2$
1	0.098306	0.0637076	0.0345984	0.026694