Projected Score Methods for Nuisance Parameters: Asymptotics and Neyman-Scott Problems

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.

Richard P. Waterman is Assistant Professor, Department of Statistics, University of Pennsylvania, Philadelphia PA 19104; and Bruce G.Lindsay is Distinguished Professor of Statistics, Department of Statistics, Pennsylvania State University, University Park, PA 16802.

The paper addresses the issue of efficient estimation of a parameter of interest as the number of nuisance parameters increases without bound. An asymptotic viewpoint is taken, termed the "rectangular array", in which the information associated with each nuisance parameter grows at the same rate as the number of nuisance parameters. It is shown that classical maximum likelihood estimation is inefficient in this framework because of a bias term in the limiting distribution of the estimate. However, by using a conceptually simple adjustment to the maximum likelihood estimating equations, an estimator can be constructed that is asymptotically unbiased and has variance that attains the Cramer–Rao lower bound.

The adjustment to the maximum likelihood estimating equations is obtained by projecting the score function for the parameter of interest onto a space that contains estimating functions that are characterized by being less sensitive to the nuisance parameters than the score equations. The value of the results show that the Cramer-Rao lower bound is still meaningful in this asymptotics, that one can construct an estimator that has asymptotic variance which attains this bound, and further that consideration of different asymptotics offers the analyst an increased choice of estimators. Natural applications are in models that have a "strata" structure, in which each strata has an associated nuisance parameter, for example the Rasch model. The adjusted maximum likelihood estimator can be viewed as an approximation to the conditional maximum likelihood estimator so that the approach described in this paper provides a bridge between conditional and unconditional inference.

Key words: Cramer–Rao lower bound; Nuisance parameters; Rectangular array.

1 INTRODUCTION

In this paper the problem of estimating a parameter of interest in the presence of nuisance parameters will be discussed. The main objective is to examine a version of the classic Neyman-Scott problem (Neyman & Scott, 1948), demonstrate the failure of maximum likelihood estimation, and then indicate how maximum likelihood estimation can be repaired through a simple adjustment to the score function for the parameter of interest.

The Neyman-Scott problem we consider is as follows. An array of independent random variables is observed:

$$\begin{matrix} X_{11} & \dots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{p1} & \dots & X_{pn} \end{matrix}$$

Row i comprises a vector of independent and identically distributed observations from common density $f(x_{ij}; \psi, \lambda_i)$, where λ_i is a row specific parameter, possibly vector valued, such as might occur when the rows correspond to a stratum or an experimental block. The vector parameter ψ corresponds to a global parameter of interest, for example treatment effect parameters.

The development of asymptotics in this problem has three natural formulations which will be referred to as *classical*, *Neyman-Scott* and *rectangular array*. The classical asymptotic formulation holds p fixed and allows n to approach infinity. Neyman-Scott asymptotics holds n fixed and p approaches infinity whilst in the rectangular array, a natural intermediate asymptotic embedding, both p and n approach infinity in such a way that their ratio approaches a non-zero constant, say $p/n \to c$, with $0 < c < \infty$. For simplicity of presentation we will sometimes focus on the case where $c = 1$, which is termed the "square array".

In the classical asymptotic setting the basic result for maximum likelihood estimation states that

$$\sqrt{np}\,(\hat{\psi} - \psi) \xrightarrow{\ell} N(0, i^{\psi\psi})\text{as } n \to \infty,$$

where $i^{\psi\psi}$ is the inverse of the average (over the strata) per observation information matrix for ψ. The maximum likelihood estimator is both consistent and efficient in the sense that its asymptotic variance is equal to the Cramér-Rao lower bound.

The fundamental results of Neyman & Scott (1948) show that if n is held fixed then the maximum likelihood estimator $\hat{\psi}$ need not be consistent as p tends to infinity, and even when $\hat{\psi}$ is consistent it may still fail to be efficient.

This paper focuses on the consequences of using likelihood methods in the rectangular array setting, which is a valuable framework to consider because it is here that maximum likelihood estimation first founders due to a bias term in the limiting normal distribution of the maximum likelihood estimate. It will be shown that a Neyman-Scott type result concerning potential inefficiency can be overcome by using a simple correction to the likelihood score equations. Part of the insight that this result provides is that it indicates that the Cramér-Rao lower bound is a meaningful lower bound in this particular asymptotic embedding even though the number of nuisance parameters increases to infinity.

Related results include those of Andersen (1970b) who showed that the Cramér-Rao lower bound is indeed a lower bound for the variance of asymptotically normal estimators in the Neyman-Scott asymptotics. However this bound is not tight in the sense that it is not attainable in general (Pfanzagl, 1993, p.1665). This is to be contrasted with the above results concerning the attainment of the lower bound in the rectangular array asymptotics.

Other results for estimation in asymptotic embeddings in which the number of parameters approach infinity include Andersen (1970a, 1970b) which contain results concerning conditional maximum likelihood estimation and Haberman (1977) which contains results for the existence and consistency of maximum likelihood estimates in exponential response models. Tak (1982) considers estimation in the classic Neyman-Scott problem and provides conditions for the consistency and asymptotic normality of the maximum likelihood estimator. Portnoy (1984, 1988) and related work concerns estimation for regression problems where the number of regression coefficients becomes large.

2 EXAMPLES

In order to illustrate the main points of the paper two examples are presented below.

The first example is the canonical Neyman-Scott problem for the inconsistency of the maximum likelihood estimator.

Example 1: Oneway analysis of variance

The random variables X_{ij} follow a normal law with mean μ_i and common variance σ^2, the variance being the parameter of interest. Regardless of the asymptotic embedding, the maximum likelihood estimator of σ^2, is $\hat{\sigma}^2 = (1/np)\sum_{ij}(x_{ij} - \overline{x}_i)^2$, but the limit in probability of this estimate depends on the asymptotic formulation. In both the classical and rectangular array settings $\hat{\sigma}^2$ tends in probability to σ^2 so that the maximum likelihood estimator is consistent. However, in the Neyman-Scott setting the limiting constant is $\{(n-1)/n\}\sigma^2$ so that the maximum likelihood estimator is inconsistent. For small n, for example when $n = 2$, this result is potentially disastrous if ignored.

In the classical asymptotics $\hat{\sigma}^2$ is efficient as well as consistent, by which is meant $\sqrt{np}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{\ell} N(0, i^{\sigma^2\sigma^2})$, as $n \to \infty$, where $i^{\sigma^2\sigma^2} = 2\sigma^4$ is the inverse of the Fisher's information about σ^2 in the presence of μ, $(i_{\sigma^2\sigma^2} - i_{\sigma^2\mu} i_{\mu\mu}^{-1} i_{\mu\sigma^2})^{-1}$ and $i_{\sigma^2\sigma^2}$ etc are the components of the Fisher's information matrix. In Neyman-Scott asymptotics the maximum likelihood estimator fails to be consistent so that it cannot possibly be efficient. In the rectangular array, it can be shown that the limiting distribution of $\sqrt{np}(\hat{\sigma}^2 - \sigma^2)$ is $N(-\sqrt{c}\sigma^2, 2\sigma^4)$ as $(n,p) \to \infty$, so that the limiting bias term, $-\sqrt{c}\sigma^2$, prohibits $\hat{\sigma}^2$ from being efficient even though $\hat{\sigma}^2$ has the correct limiting variance.

Example 2: Circular normal

Pairs of independent normal random variables are observed with distributions

$$X_{i1} \sim N(\rho \cos \lambda_i, 1) \qquad X_{i2} \sim N(\rho \sin \lambda_i, 1).$$

This experiment corresponds to observing a series of bivariate normal random variables with means lying on a circle of unknown radius ρ. The radius is the parameter of interest and the nuisance parameters are the angles λ_i. The value of this example lies in the fact that there is natural curvature to the problem which plays a fundamental role in the more general setting. There is no closed form expression for the maximum likelihood estimator in this model, but as n becomes large for p fixed at $p = 1$, then for $\rho > 1$ the maximum likelihood estimator of ρ tends in probability to a constant that is well approximated by $\rho + 1/(2\rho)$, so that the maximum likelihood estimator is again inconsistent in the Neyman-Scott framework.

3 TOOLS

We assume the following distributional setting: let Ψ and Λ be two subsets of R, with $\psi \in \Psi$ and $\lambda_i \in \Lambda$. Let $\mathcal{X}$ be the common sample space for $X_{ij}, i =$

$1, \cdots, p, \quad j = 1, \cdots, n$. Let $\mathcal{B}$ be a σ-field over $\mathcal{X}$ and $\mathcal{P} = \{P(\,.\,|\psi, \lambda) : \psi \in \Psi, \lambda \in \Lambda\}$ be a family of probability measures defined on $\mathcal{B}$. Suppose that there exists a σ-finite measure μ on $\mathcal{B}$ such that $\mathcal{P}$ is dominated by μ. Let $f(x|\psi, \lambda)$ be the density $dP(x|\psi, \lambda)/d\mu(x)$.

This section will concentrate on a single strata of the rectangular array. In §5 the original rectangular array structure will be recovered by appealing to a simple additivity result that exploits the independence of observations between the strata. In a single strata there is only one nuisance parameter, the index i is fixed and is therefore omitted.

The technique used to repair the maximum likelihood estimating equations is the projected score approach. Details can be found in Small & McLeish (1989) and Waterman & Lindsay (1996a). The aim of the procedure is to reduce the impact of the nuisance parameters on the estimating equation for ψ. This is achieved by removing from U_0 its projection onto the space $\mathcal{V}_t$, where $\mathcal{V}_t$ is the closed linear space spanned by the functions $V_k = f^{(k)}/f$ for $k = 1, \cdots, t$. Differentiation is with respect to the nuisance parameter. This basis is referred to as the Bhattacharyya basis.

Denoting the projection operator onto $\mathcal{V}_t$ by Π_t the projected scores are defined as

$$U_t = U_0 - \Pi_t U_0.$$

In the case where a complete and sufficient statistic, T, exists for λ then the space spanned by $\mathcal{V}_t$ as $t \to \infty$ is the space of T-measurable functions, so that $\Pi_t U_0 = E(U_0 \,|\, T)$ and U_t approaches the conditional score function. This is a strong motivation for using the projections since they have this interpretation as approximations to the conditional score function, and the optimality properties of the conditional score function are well known (Andersen, 1970a; Godambe, 1976; Lindsay, 1982).

Simplification can be achieved by using an orthogonalized version of the Bhattacharyya basis. Elements of the orthogonalized basis are denoted by a $*$ superscript. Using this notation the projected scores can be written as

$$\begin{aligned} U_1 &= U_0 - \rho_{11} V_1^* \\ U_2 &= U_0 - \rho_{21} V_1^* - \rho_{22} V_2^* \\ U_t &= U_0 - \sum_{i=1}^{t} \rho_{ti} V_i^*. \end{aligned}$$

The ρ coefficients are regression coefficients defined through the relevant orthogonality constraint, ie ρ_{11} solves the equation $E\{(U_0 - \rho_{11} V_1^*)V_1\} = 0$.

Following Godambe (1976) the information in the function g is defined as

$$I_g = E(\nabla g)\{E(gg^T)\}^{-1} E(\nabla g^T),$$

so that the information in U_t, written as I_t can be expressed as

$$I_t = I_{t-1} - E(U_0 V_t^*) E(V_t^* V_t^*)^{-1} E(V_t^* U_0),$$

indicating that the information is decreasing in t. The order of magnitude of the information change $I_t - I_{t-1}$ is $O(n^{2-t})$, where n is the stratum sample size. The information change has the potential to decrease very quickly in t. Note that I_1 is the marginal Fisher's information for ψ. Building from these within strata projected score functions this paper considers the derived U_1 and U_2 projected score functions for the rectangular array. The function U_1 provides maximum likelihood estimates, denoted by $\hat{\psi}$ and U_2 provides adjusted maximum likelihood estimates denoted by $\tilde{\psi}$. The main aim here is to show that $\tilde{\psi}$ is a fully efficient estimate in the rectangular array.

4 LIMITING DISTRIBUTIONS IN THE STRATA

The estimation of the nuisance parameters has a cost. In this section we develop approximations for the cost in bias and information loss based on using second order asymptotics as $n \to \infty$ within a single stratum. The asymptotic distribution results are obtained using standard Taylor series arguments. The results have multivariate analogs, but only the univariate ψ case will be considered here. Two results are used in the proofs: the first of these is a U-statistic representation theorem for the Bhattacharyya basis functions, and the second is an expectation identity type result similar in nature to the Bartlett identities. In the univariate case the identity states that for a function T of the data and parameters, satisfying smoothness conditions that allow interchange of the expectation and differentiation operations, which in addition is orthogonal to the Bhattacharyya basis of order $t-1$:

$$E(T\,V_t) = (-1)^{t-k} E(T^{(k)} V_{t-k}), \text{ for } k = 0, \dots t. \tag{1}$$

For example, with $T = V_1$ and $t = 1$ we obtain the familiar information identity $E(V_1^2) = -E(V_1')$, and for $T = V_2^*$, and $t = 2$ we obtain $E(V_2^{*2}) = -E(V_2^{*\prime} V_1) = E(V_2^{*\prime\prime})$. The usefulness is that expectations involving derivatives of a function are expressed in terms of covariances. See Waterman & Lindsay (1996a) for more discussion of these results.

The claim that the projective scores "remove bias" is partially illustrated through the following results which describe precisely the manner in which estimating the nuisance parameters with their maximum likelihood estimate, $\hat\lambda_\psi$, induces bias and information loss in the score functions.

$$U_1(\psi, \hat\lambda_\psi) - U_1(\psi, \lambda) \xrightarrow{\ell} Y_1 + \tfrac{1}{2}\kappa \tag{2}$$

$$U_2(\psi, \hat\lambda_\psi) - U_1(\psi, \lambda) \xrightarrow{\ell} Y_1 \tag{3}$$

$$U_2(\psi, \hat\lambda_\psi) - U_2(\psi, \lambda) \xrightarrow{\ell} Y_2 \tag{4}$$

Both Y_1 and Y_2 are mean zero quadratic forms in normal random variables. The equations have the following interpretations: equation (2) shows the inherent bias in the estimated U_1 equation which leads to the asymptotic bias of the maximum likelihood estimator in the rectangular array setting. Equation (3) indicates that the estimated U_2 removes this bias without increasing variance so that the estimated U_2 score is closer to the actual U_1 score than is the estimated U_1 score. Finally (4) indicates that the estimated U_2 is closer to the true U_2 than it is to the true U_1 score because the variance of Y_2 is less than that of Y_1. The constant κ has an interpretation as a model curvature, and so is invariant under interest preserving transformations. In Example 2, the circular normal problem, this curvature is precisely that of the circle on which the mean lies, so that the tighter the circle the larger the bias and the greater the decrease in information between U_1 and U_2. The critical result here is that the limiting distribution of $\hat{U}_1 - U_1$ contains a bias term and this bias accumulates in the rectangular array setting leading to inefficiency of the maximum likelihood estimator.

5 APPLICATION TO THE NEYMAN–SCOTT PROBLEM

In this section the main result of the paper is presented.

The following lemma presents the additivity result that allows the projected scores to be reconstructed from the within strata projected scores, the function U_{ti} is the t-th order projected score from the i-th stratum.

Lemma 5.1: In an independent strata sampling scheme

$$U_t = \sum_{i=1}^{p} U_{ti}, \text{ and the information } I_{U_t} = \sum_{i=1}^{p} E(U_{ti} U_{ti}^T). \tag{5}$$

The lemma states that both the projective scores and the information are additive over the strata. The proof is straightforward and is omitted.

Before proceeding to a statement of the theorem we will briefly discuss some regularity conditions:

- **Smoothness:** we assume that the likelihood is sufficiently smooth so that all necessary higher order derivatives exist and are square integrable with respect to the dominating measure μ. Differentiation and integration operators may be exchanged so that for example, the score functions are "information unbiased"; $-E(\partial U_1/\partial\psi) = E(U_1^2)$.

- **Asymptotic negligibility of the remainder term:** the proof uses a second order von-Mises expansion. The expansion will contain a remainder term which will be assumed to be $o_p(1)$ with respect to the main terms. This is the analogy of the usual bounded third derivative assumption for second order expansions.

- **Regularity of the nuisance parameters:** We take the most simple, and not unrealistic approach of assuming that Λ is a closed interval in R.

- **Uniformity of convergence over the strata:** limiting results will be applied within each strata. For example, standardized within strata score functions can be expected to converge to normal random variables; we assume that this convergence is uniform across the strata.

We consider $\hat{\psi}$, $\tilde{\psi}$ and $\hat{\lambda}_{i\psi}$ as functionals on $F = (F_1, \ldots, F_p)$ where F_i is the stratum distribution function. In each stratum consider a perturbed version of F_i given by $F_i(\epsilon_i) = F_i + \epsilon_i\sqrt{n}(\hat{F}_i - F_i)$ where n is the sample size in the i^{th} stratum and $\hat{F}_i$ the empirical distribution function. Define $\boldsymbol{\epsilon}$ to be the vector $(\epsilon_1, \ldots, \epsilon_p)$. The maximum likelihood estimator for λ_i at a fixed ψ is written as $\lambda_i\{\psi(F(\boldsymbol{\epsilon})), F_i(\epsilon_i)\}$. In what follows, functions with triple subscripts, for example U_{1ij}, denote the score functions based on the likelihood of the j-th observation in the i-th stratum.

Theorem 5.1: Under appropriate regularity conditions in the rectangular array embedding ($\lim_{p,n\to\infty} \frac{p}{n} = c > 0$) the maximum likelihood solution $\hat{\psi} = \psi(\hat{F})$ to the overall score equation $U_1 = 0$ has limiting distribution

$$\sqrt{np\overline{I}_p}\left(\psi(\hat{F}) - \psi(F_0)\right) \xrightarrow{\ell} N(b,1).$$

The adjusted estimate, $\tilde{\psi} = \psi(\hat{F})$, the solution to overall score equation $U_2 = 0$, has limiting distribution

$$\sqrt{np\overline{I}_p}\left(\psi(\hat{F}) - \psi(F_0)\right) \xrightarrow{\ell} N(0,1).$$

The limiting bias b is equal to

$$b = -\frac{1}{2}\sqrt{c\overline{I}}\lim_{p\to\infty}\sum_{k=1}^{p}\frac{E(U_{1k1}V_{2k1}^{*})/E(V_{1k1}^{2})}{p\overline{I}_p}. \tag{6}$$

Here, $\overline{I} = \lim_{p\to\infty}\overline{I}_p$ where $\overline{I}_p = \sum_{i=1}^{p} E(U_{1i1}^2)/p$, and $0 < \overline{I} < \infty$.

Note that $\hat{\psi}$ has a non-zero mean limiting distribution and that $\tilde{\psi}$ has a mean zero limiting distribution but that both have the same asymptotic variance.

A description of the proof follows; further details are in the appendix. The von-Mises expansion of $\psi(F)$ about $\psi(F_0)$, evaluated at $\hat{F}$ is defined as

$$\psi(\hat{F})-\psi(F_0) = \sum_{i=1}^{p}\frac{1}{\sqrt{n}}\left.\frac{d\psi(F(\boldsymbol{\epsilon}))}{d\epsilon_i}\right|_{\epsilon_i=0} + \frac{1}{2}\sum_{i,j}\frac{1}{\sqrt{n}\sqrt{n}}\left.\frac{d^2\psi(F(\boldsymbol{\epsilon}))}{d\epsilon_i d\epsilon_j}\right|_{\epsilon_i=\epsilon_j=0} + Rem(\hat{F}-F_0). \tag{7}$$

By rewriting the estimating equations U_1 and U_2 in functional form (see (A·5) in the appendix) and differentiating with respect to $\boldsymbol{\epsilon}$ it is possible to implicitly calculate the derivatives

$$\frac{d\psi(F(\boldsymbol{\epsilon}))}{d\epsilon_i} \qquad \text{and} \qquad \frac{d^2\psi(F(\boldsymbol{\epsilon}))}{d\epsilon_i d\epsilon_j},$$

and hence to examine the asymptotic distributional form of (7). After normalizing by the appropriate constant, $\sqrt{np\overline{I}_p}$, the first term on the RHS of (7) can be shown to have an $N(0,1)$ limit, whereas the second term goes in probability to a constant. This constant is b when U_1 is used for estimation but 0 if U_2 is used.

As an illustration of the bias calculation we derive the relevant terms for Example 1. In this case all the elements of the bias term b, are free of the nuisance parameter, so that the terms in the summations within b are identical and therefore only need to be calculated for a single strata. The elements of b can be extracted from the within strata *extended information matrix* for a single observation, the matrix being defined as $M_2 = E\{(U_{0i1}, V_{1i1}, V_{2i1})^T\,(U_{0i1}, V_{1i1}, V_{2i1})\}$.

$$M_2 = \begin{pmatrix} \frac{1}{2\sigma^4} & 0 & \frac{1}{\sigma^4} \\ 0 & \frac{1}{\sigma^2} & 0 \\ \frac{1}{\sigma^4} & 0 & \frac{2}{\sigma^4} \end{pmatrix}.$$

One can verify that identifying the terms in M_2 with the elements of b given in (6) does indeed give the bias of the maximum likelihood estimator in the limiting distribution as $-\sqrt{c}\sigma^2$. To be specific, $E(U_{1i1}V^*_{2i1}) = 1/\sigma^4$, $E(V^2_{1i1}) = 1/\sigma^2$ and $\overline{I}_p = 1/(2\sigma^2)$.

6 NUMERICAL RESULTS AND A SIMULATION

Here we present the results of some simulations that serve to verify the problem that bias in the limiting distribution of the maximum likelihood estimator induces on confidence intervals for the parameter of interest.

We start with Example 1, the standard Neyman–Scott problem. Exact coverage rates were calculated for the following three asymptotic pivotal quantities. The first pivotal is for the maximum likelihood estimator the second for $\tilde{\psi}$ the root of U_2 which is in fact the conditional maximum likelihood estimator in this problem (see Waterman & Lindsay (1996b) for more details on the links between conditional and projected inference), and the third pivotal is for an information corrected version of U_2 which we briefly discuss and is the subject of further research. Exact 95% coverage rates were calculated using the numerical integration function `NIntegrate` function in Mathematica. For convenience all the nuisance parameters were set to $\lambda_i = 0$, and σ^2 equal to 1. The asymptotic pivotals are

$$\begin{array}{ccc} 1 & 2 & 3 \\ \frac{\sqrt{n\,p}(\hat{\psi}-\psi)}{\sqrt{2}\psi} & \frac{\sqrt{n\,p}(\tilde{\psi}-\psi)}{\sqrt{2}\psi} & a(\kappa,i)\frac{\sqrt{n\,p}(\tilde{\psi}-\psi)}{\sqrt{2}\psi} \end{array}$$

where $a(\kappa, i) = 1 - \frac{1}{2\,n}$.

INSERT TABLE 1 ABOUT HERE

Note that the maximum likelihood estimator coverage rates, in column 1 are systematically biased, whereas the U_2 coverage rates, in column 2, approach the correct nominal value of 0.95. However, the extremely good properties of the information corrected estimate obviously warrant further investigation.

As a final example we present an extension of the binary matched pairs experiment. Let X_i and $Y_i, i = 1, \cdots, p$ be independent binomial random variables based on $n = p$ trials and with success probabilities of $q(0, \lambda_i)$ and $q(\psi, \lambda_i)$ respectively. We take q^{-1} to be the complementary log-log link, so that

$$q(\psi, \lambda_i) = 1 - \exp\{-\exp(\psi + \lambda_i)\}.$$

This example illustrates the value of the projected score method because the parameter of interest is not a canonical parameter so that conditioning arguments for the removal of λ_i do not apply. Use of the projected scores is made feasible through the aid of a symbolic math package. This is useful in calculating terms such as $E(V_{2ij}V_{2ij})$, which is considerably more complicated in this example than in the previous one. The simulation, carried out in `SPlus` consisted of performing 1000 replicates at each level of p with the number of strata ranging from 10 to 20. The nuisance parameters were set equal at 0.0, whilst the parameter of interest was set at 0.75. Both the maximum likelihood estimator $\hat{\psi}$ and the adjusted maximum likelihood estimator $\tilde{\psi}$ were calculated along with their estimated standard errors. The results reported below show the proportion of times the 90% confidence interval, constructed as $\hat{\psi} \pm 1.645\widehat{s.e.}(\hat{\psi})$ covered the true value, and likewise for the adjusted maximum likelihood estimator.

INSERT TABLE 2 ABOUT HERE

The table again corroborates the assertion of the main theorem, with the maximum likelihood estimator performing badly, but the corrected maximum likelihood estimator performing well. The difference in coverage rates is a function of the curvature κ, so that in these simulations the extent of the failure of the maximum likelihood estimator can be made large or small simply by choosing the parameter values carefully. The reason for not including the information corrected pivotal is that it is considerably more complicated than in Example 1 because of the absence of a complete and sufficient statistic for the nuisance parameters.

7 DISCUSSION

In the paper we have only considered balanced within strata sample sizes. The results still hold in the case of unbalanced strata sample sizes so long as no strata dominates in the limit, that is each of the summands in $\sum_i E(U_{1i1}^2)$

should be asymptotically negligible. The situation will change however if the within strata sample size grows at a different rate than the number of strata, that is $p = n^{1+\alpha}$, for example if $\alpha = 1$ so that $n = \sqrt{p}$. In this particular situation we conjecture that the higher order projected scores, for example U_3, will be needed to remove bias from the limiting distribution of the maximum likelihood estimator.

The paper concentrated on univariate ψ and λ_i but this was only for convenience. One of the strengths of the methodology is that it has a straightforward extension to the multivariate case. Future work in this area concerns a characterization of the second order information loss in the infinite parameter setting developing the information corrections as reported in Table 1.

A APPENDIX: PROOF OF THE MAIN THEOREM

The proof and its components have been arranged into the following sections: the first section addresses the underlying assumptions, the second presents the second order von-Mises expansion for functionals, the third derives relevant properties of the functional $\lambda_i(\psi, F_i)$ the maximum likelihood estimator of λ_i for ψ fixed. The fourth section presents the functional representation of the estimating functions U_1 and U_2, and the fifth and sixth sections evaluate the stochastic properties of the first and second order derivatives in the von-Mises expansion that are required in order to derive the limiting distribution of the functionals themselves.

Assumptions

Distributional results are obtained in the proof by appealing to the Central Limit Theorem for triangular arrays (Billingsley, 1986, Theorem 27.2) so that the essential assumptions are those that establish its applicability.

To formulate the assumptions define the triangular array with the p-th row having elements

$$A_{ni} = \sqrt{n}\frac{\sum_{j=1}^{n} U_{1ij}}{n}, i = 1, \cdots, p_n.$$

Note that p is a function of n, in particular in the square array $p = n$ but in what follows we write p rather than p_n for brevity. Let $\sigma_{ni}^2 = E(A_{ni}^2) = E(U_{1i1}^2)$ and $s_n^2 = \sum_{i=1}^{p} \sigma_{ni}^2 = I_{U_1}$.

Then provided that the Lindeberg condition is satisfied $\sum_{i=1}^{p} A_{ni}/s_n \xrightarrow{\ell} N(0, 1)$ as $n \to \infty$, so we assume that $0 < E(U_{1i1}^2) < \infty$ for all i, and that $0 < \lim_{p\to\infty} \overline{I}_p < \infty$.

As $n \to \infty$ the standardized row score variables A_{ni} approach a normal limit, denoted by Y_i. This follows from the regular central limit theorem for independently identically distributed summands. This motivates the following assumptions which are required to hold for all sequences $\{\lambda_p\} \in \Lambda$.

- (A.1) Uniformly over the strata $E(A_{ni}^4) \to E(Y_i^4)$.
- (A.2) There exist constants $B > 0$ and $C < \infty$, and an integer N_0 such that

$$\text{Var}(A_{ni}) \geq B \quad \text{for all i, and } n \geq N_0$$

and

$$E(A_{ni}^4) \leq C \quad \text{for all i, and } n \geq N_0$$

These moment assumptions are enough to ensure that Lyapounov's condition with $\delta = 2$ holds, from which Lindenberg's condition follows so that the triangular array central limit theorem applies to the variables A_{ni}.

We assume a similar set of conditions so that the triangular array central limit theorem applies to the variables $\lambda_i^{\epsilon_i}$ to be defined defined in (A·4).

Basic set up

For the i^{th} stratum consider the perturbation in the distribution function F_{i0} in the direction $\hat{F}_i - F_{i0}$ given by

$$F_i(\epsilon_i) = F_{i0} + \epsilon_i\sqrt{n}(\hat{F}_i - F_{i0}). \tag{A·1}$$

The derivative of (A·1) with respect to ϵ_i is

$$\sqrt{n}(\hat{F}_i - F_i) = \Delta_{in}.$$

Note that if F_i is evaluated at $\epsilon_i = 0$ and $\epsilon_i = \frac{1}{\sqrt{n}}$ then we obtain

$$F_i(0) = F_{i0} \qquad \text{and} \qquad F_i\left(\frac{1}{\sqrt{n}}\right) = \hat{F}_i.$$

The second order von-Mises expansion is given by

$$\psi\left(\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}}\right)-\psi(0,\ldots,0)=\sum_k \psi'_k\Big|_{\epsilon_k=0}\frac{1}{\sqrt{n}}+\frac{1}{2}\sum_{kl}\psi''_{kl}\Big|_{\epsilon_k=\epsilon_l=0}\frac{1}{\sqrt{n}}\frac{1}{\sqrt{n}}+Rem(\hat{F}-F_0),$$

where ψ'_k and ψ''_{kl} are defined as in v. Mises (1947, p.323, p.324).

Limiting distributions will be derived by considering the quantity

$$\sqrt{n\,p\,\overline{I}_p}\left(\psi\left(\frac{1}{\sqrt{n}},\cdots,\frac{1}{\sqrt{n}}\right)-\psi(0,\cdots,0)\right)-\sqrt{n\,p\,\overline{I}_p}Rem(\hat{F}-F_0) \;=\; \sqrt{\overline{I}_p}\left(\sum_k \sqrt{p}\psi'_k+\frac{1}{2}\sum_{kl}\psi''_{kl}\right) \qquad (\text{A}\cdot 2)$$

under the assumption that $\sqrt{n\,p\,\overline{I}_p}Rem(\hat{F}-F_0)=o_p(1)$. The proof proceeds by examining the limiting distributions of the two terms on the RHS of (A·2). The first term is shown to have a standard normal limit regardless of which of U_1 or U_2 is used for estimation whereas the second term is shown to converge in probability to 0 if U_2 is used but to a non-zero constant if U_1 is used.

Distributional properties of $\lambda_i(\psi, F_i)$

This section provides some facts about the functional $\lambda_i(\psi, F_i)$ the solution to the maximum likelihood score equation for the nuisance parameter when the parameter of interest ψ is fixed. These facts are needed because the $\lambda_i(\psi, F_i)$ are the plug-in values that are used to estimate the nuisance parameters.

In the i^{th} stratum $\lambda_i(\psi, F_i(\epsilon_i))$ solves

$$\int V_{1ij}[\psi,\lambda_i\{\psi,F_i(\epsilon_i)\}]dF_i(\epsilon_i)=0, \qquad (\text{A}\cdot 3)$$

where the function V_{1ij} is the nuisance parameter score for the jth observation in stratum i. Equation (A·3) is a function of two variables, ψ and ϵ_i, and has ψ and ϵ_i-derivatives of

$$\int V^{\psi}_{1ij}[\psi,\lambda_i\{\psi,F_i(\epsilon_i)\}]dF_i(\epsilon_i)+\lambda^{\psi}_i\{\psi,F_i(\epsilon_i)\}\int V^{\lambda_i}_{1ij}[\psi,\lambda_i\{\psi,F_i(\epsilon_i)\}]dF_i(\epsilon_i),$$

and

$$\lambda_i^{\epsilon_i}\{\psi, F_i(\epsilon_i)\}\int V_{1ij}^{\lambda_i}[\psi,\lambda_i\{\psi,F_i(\epsilon_i)\}]dF_i(\epsilon_i)+\int V_{1ij}[\psi,\lambda_i\{\psi,F_i(\epsilon_i)\}]d\Delta_{in},$$

respectively. Equating these equations to zero and solving for the derivatives of λ_i evaluated at $\epsilon_i = 0$ gives

$$\lambda_i^{\psi} = -\frac{E(V_{1i1}^{\psi})}{E(V_{1i1}^{\lambda})},$$

and

$$\lambda_i^{\epsilon_i} = -\frac{\int V_{1ij}d\Delta_{in}}{\int V_{1ij}^{\lambda_i}dF_i} = \sqrt{n}\frac{\sum_j V_{1ij}}{n\,E(V_{1ij}^2)}. \tag{A·4}$$

In this expansion λ_i^{ψ} is $O(1)$ but $\lambda_i^{\epsilon_i}$ is an $O_p(1)$ random variable. We have used the fact that $E(V_{1i1}^{\lambda_i}) = -E(V_{1i1}^2)$ to simplify the denominator in (A·4).

Expressing U_{2i} in functional form

In the ith stratum the second order score equation U_{2i} can be written as $U_{2i} = U_{1i} - \rho_i V_{2i}^*$, where V_{2i}^* is the component of V_{2i} orthogonal to V_{1i}, and ρ_i is the appropriate regression coefficient. By definition $V_{2i}^* = V_{2i} - \tau_i V_{1i}$ where again τ_i makes the appropriate orthogonality adjustment. Now ρ_i by definition equals $\frac{E(U_{1i}V_{2i}^*)}{E({V_{2i}^*}^2)}$, which can be written as

$$\frac{1}{n}\frac{E(U_{1i1}V_{2i1}^*)}{2\{E(V_{1i1}^2)\}^2} + O\left(\frac{1}{n^2}\right).$$

Likewise the term τ_i can be shown to be $O(\frac{1}{n})$.

Consequently the U_{2i} equation can be written in functional form as

$$\begin{aligned} n^{-1}U_{2i} &= \int U_{1ij}dF_i(\epsilon_i) - \tilde{\rho}_i\left[\left(\int V_{1ij}dF_i(\epsilon_i)\right)^2\right. \\ &+ \left.\epsilon_i^2\left(\int V_{1ij}^{\lambda_i}dF_i(\epsilon_i)\right) + \epsilon_i^2\tau_{i1}\left(\int V_{1ij}dF_i(\epsilon_i)\right)^2 + \epsilon_i^4\tau_{i2}\int W_{1ij}dF_i(\epsilon_i)\right], \end{aligned} \tag{A·5}$$

where the terms involving n have been taken out of the regression coefficients and replaced by the ϵ_i terms, so that for example $\tilde{\rho}_i = E(U_{1i1}V^*_{2i1})/(2\{E(V^2_{1i1})\}^2)$. The components τ_{i1} and τ_{i2} contain various products of the regression coefficients and the term W_{1ij} is a weighted sum of V_{1ij} and $V^{\lambda_i}_{1ij}$. Evaluating (A·5) at $\epsilon_i = 1/\sqrt{n}$ returns U_{2i}/n.

We will differentiate (A·5) repeatedly and make the following remarks to aid simplification:

1 Since estimating functions are to be evaluated at the maximum likelihood estimator for the nuisance parameter, any term that involves the integration of V_{1ij} with respect to $dF_i(\epsilon_i)$ will be identically zero; see (A·3).

2 The term involving ϵ_i^4 would require a fourth order derivative to be taken before it would contribute to the asymptotic analysis as ϵ_i is set equal to 0. Since only second derivatives are used in the expansion this term is redundant at this level of expansion.

3 We split the whole $\boldsymbol{\epsilon}$ into components ($\boldsymbol{\epsilon}_1, \epsilon_2, \epsilon_3$ and ϵ_4) to aid bookkeeping and to identify easily those differentiations that will be redundant. To this end the following form of (A·5) is used:

$$U_2^\dagger(\boldsymbol{\epsilon}) = \sum_i \Big\{ \int U_{1ij}(\psi(\boldsymbol{\epsilon}_1), \lambda_i(\psi(\boldsymbol{\epsilon}_1), \epsilon_{2i}))dF_i(\epsilon_{3i}) - \underbrace{\rho_i(\psi(\boldsymbol{\epsilon}_1), \lambda_i(\psi(\boldsymbol{\epsilon}_1), \epsilon_{2i})) \left(\epsilon_{4i}^2 \int V^{\lambda_i}_{1ij}(\psi(\boldsymbol{\epsilon}_1), \lambda_i(\psi(\boldsymbol{\epsilon}_1), \epsilon_{2i}))dF_i(\epsilon_{3i})\right)\Big\}}_{\text{Correction term}} . \qquad \text{(A·6)}$$

The first term U_{1ij} could be written as U_{0ij} since we are evaluating at $\lambda_i(\psi, F_i(\epsilon_i))$, but it is more convenient to leave it in this form as terms such as $U^{\lambda_i}_{1ij}$ are mean zero. One of the above ϵ's is a vector, $\boldsymbol{\epsilon}_1$, the rest are scalars, $\epsilon_{2i}, \epsilon_{3i}$ and ϵ_{4i}.

First order differentiation

Note that in the first order differentiation the correction term in (A·6) will not contribute since all ϵ's will be set equal to zero for evaluation of derivatives.

Since

$$\frac{d}{d\epsilon_k} = \frac{d}{d\epsilon_{1k}} + \frac{d}{d\epsilon_{2k}} + \frac{d}{d\epsilon_{3k}} + \frac{d}{d\epsilon_{4k}},$$

we have that

$$\begin{aligned} \frac{d}{d\epsilon_k} U_2(\boldsymbol{\epsilon}) \;&=\; \sum_i \left\{ \int \left(U_{1ij}^{\psi} \psi^{\epsilon_{1k}} + U_{1ij}^{\lambda_i} \psi^{\epsilon_{1k}} \lambda_i^{\psi} \right) dF_i(\epsilon_{3i}) \right\} + && \text{(A·7)} \\ & \int U_{1kj}^{\lambda_k} \lambda_k^{\epsilon_{2k}} dF_k(\epsilon_{3k}) + && \text{(A·8)} \\ & \int U_{1kj} d\Delta_{kn} + && \text{(A·9)} \\ & \text{derivative of correction term.} \end{aligned}$$

The basic heuristic for simplification is that a mean zero function inside an integral which has not been differentiated with respect to ϵ will be equal to 0 on evaluation at $\epsilon = 0$. Also, terms such as $\psi^{\epsilon_{1k}}$ and $\lambda_k^{\epsilon_{2k}}$ can be passed through the integral. On setting $dU_2/d\epsilon_k$ equal to zero and evaluating at $\epsilon = 0$ we obtain

$$\psi'_k \sum_i E(U_{1i1}^{\psi}) + \sqrt{n}\frac{\sum_j U_{1kj}}{n} = 0.$$

(We have written ψ'_k for $\psi^{\epsilon_{1k}}$).

Recall from (A·2) that ψ'_k will be multiplied by $\sqrt{p\,\overline{I}_p}$. After summing ψ'_k over k and performing this multiplication we obtain

$$\sqrt{p\,\overline{I}_p} \sum_k \psi'_k = -\sqrt{p\,\overline{I}_p} \frac{\sum_k \sqrt{n}\frac{\sum_j U_{1kj}}{n}}{\sum_i E(U_{1i1}^{\psi})},$$

which is $O_p(1)$ by the assumptions that ensure a triangular array central limit theorem applies for the variables $\sqrt{n}\frac{\sum_j U_{1ij}}{n}$. Note that the limiting variance is 1. This is the first order term of an expansion of the root of either U_1 or

U_2. The asymptotic behavior of the roots differs because of the differences in the second order terms that are derived in the next section.

Second order terms

The basic outline of this stage of the proof is as follows: we are interested in the second order term

$$\frac{1}{2}\sum_{kl} \psi''_{kl}\bigg|_{\epsilon_k=\epsilon_l=0} \frac{1}{\sqrt{n}}\frac{1}{\sqrt{n}}. \tag{A·10}$$

This is multiplied by $\sqrt{n\,p\,\overline{I}_p}$ to obtain its contribution to the limiting distribution which is therefore

$$\frac{1}{2}\sqrt{\overline{I}_p}\sum_{kl} \psi''_{kl}\bigg|_{\epsilon_k=\epsilon_l=0}. \tag{A·11}$$

The ψ''_{kl} will be found implicitly by differentiating (A·6) twice and equating the derivatives to zero.

Since the only difference between the maximum likelihood score equations and the $U_2^\dagger$ score equations is the correction term in (A·6) both parts of the proof of Theorem 5.1 will be done simultaneously.

The claim is that the second derivative of (A·6) with respect to ϵ_k and ϵ_l on evaluation at $\boldsymbol{\epsilon} = 0$ is of the form:

$$\sum_{kl} \frac{d^2}{d\epsilon_k d\epsilon_l} U_2^\dagger(\boldsymbol{\epsilon}) = A\,B + C + D = 0, \tag{A·12}$$

where $A = \sum_{kl} \psi''_{kl}$, $B = \sum_i E(U^{\psi}_{1i1}) = O(p)$, $C = \sum_k \tilde{U}_k$ where the $\tilde{U}_k$ are mean zero random variables when the U_2 score equation is used for estimation, but not mean zero when the mle score U_1 is used, and D comprises remainder terms of magnitude at most $O_p(1)$. By making A in (A·12) the subject of the formula and by applying a triangular array Law of Large Numbers (Billingsley, 1986, Theorem 6.2) for the sum $\sum_k \tilde{U}_k/p$, establishes that the only contribution of $\sum_{kl} \psi''_{kl}$ to the limiting distribution is a constant. The objective is to identify this constant firstly when U_1 is used for estimation and secondly when U_2 is used.

Note that the correction term in (A·6) only contributes to the limiting distribution when we differentiate twice with respect to ϵ_{4k}. The contribution is

$$-2\tilde{\rho}_k E(V_{1k1}^{\lambda_k}) = \frac{2E(U_{1k1}V_{2k1}^*)}{2\{E(V_{1k1}^2)\}^2} E(V_{1k1}^2) = \frac{E(U_{1k1}V_{2k1}^*)}{E(V_{1k1}^2)}. \tag{A·13}$$

This is a component of the term that appears in the limiting distribution of the root of U_2 but not in the distribution of the root of U_1.

Proceeding through the other second order differentiations of (A·6) we start by differentiating the term (A·7) with respect to the three working ϵ's. Differentiating with respect to ϵ_{1l} produces the term associated with the derivative of $\psi^{\epsilon_{1k}}$:

$$\psi^{\epsilon_{1k}\epsilon_{1l}} \sum_i \left\{ \int \left(U_{1ij}^{\psi} + U_{1ij}^{\lambda_i} \lambda_i^{\psi} \right) dF_i \right\}. \tag{A·14}$$

This is the only important term that comes from (A·7). The term inside the parentheses in (A·14), on evaluation at $\boldsymbol{\epsilon} = 0$ is $\sum_i E(U_{1i1}^{\psi})$, the B term in (A·12) which is $O(p)$, so it only needs to be checked in order to verify the claim of (A·12) that the other terms are at most $O_p(1)$ and are therefore subsumed into the D term. Taking the remaining derivatives of (A·7) with respect to $\epsilon_{1l}, \epsilon_{2l}$ and ϵ_{3l} gives

$$\psi^{\epsilon_{1k}} \sum_i \int \left\{ U_{1ij}^{\psi\psi} \psi^{\epsilon_{1l}} + U_{1ij}^{\psi\lambda_i} \lambda_i^{\psi} \psi^{\epsilon_{1l}} + \left(U_{1ij}^{\lambda_i\psi} \psi^{\epsilon_{1l}} + U_{1ij}^{\lambda_i\lambda_i} \lambda_i^{\psi} \psi^{\epsilon_{1l}} \right) \lambda_i^{\psi} + U_{1ij}^{\lambda_i} \lambda_i^{\psi\psi} \psi^{\epsilon_{1l}} \right\} dF_i(\epsilon_{3i}) +$$
$$\psi^{\epsilon_{1k}} \int \left\{ \left(U_{1lj}^{\psi\lambda_l} + U_{1lj}^{\lambda_l\lambda_l} \lambda_l^{\psi} \right) \lambda_l^{\epsilon_{2l}} + U_{1lj}^{\lambda_l} \lambda_l^{\psi\epsilon_{2l}} \right\} dF_l(\epsilon_{3l}) +$$
$$\psi^{\epsilon_{1k}} \int \left(U_{1lj}^{\psi} + U_{1lj}^{\lambda_l} \lambda_l^{\psi} \right) d\Delta_{ln}$$

It can be verified that on summation over k and l none of these terms has magnitude higher than $O_p(1)$.

On differentiating the term (A·8) one obtains (noting whether or not $k = l$)

$$\int \left(U_{1kj}^{\lambda_k\psi}\psi^{\epsilon_{1l}} + U_{1kj}^{\lambda_k\lambda_l}\psi^{\epsilon_{1l}}\lambda_k^{\psi}\right)\lambda_k^{\epsilon_{2k}} + U_{1kj}^{\lambda_k}\lambda_k^{\epsilon_{2k}\epsilon_{1l}}\psi^{\epsilon_{1l}}dF_k(\epsilon_{3k}) +$$
$$\int U_{1kj}^{\lambda_k\lambda_k}\lambda_k^{\epsilon_{2k}}\lambda_k^{\epsilon_{2k}} + U_{1kj}^{\lambda_k}\lambda_k^{\epsilon_{2k}\epsilon_{2k}}dF_k(\epsilon_{3k}) +$$
$$\int U_{1kj}^{\lambda_k}\lambda_k^{\epsilon_{2k}}d\Delta_{kn}$$

and finally differentiating the third term (A·9):

$$\int \left(U_{1kj}^{\psi}\psi^{\epsilon_{1l}} + U_{1kj}^{\lambda_k}\lambda_k^{\psi}\psi^{\epsilon_{1l}}\right)d\Delta_{kn} + \int U_{1k}^{\lambda_k}\lambda_k^{\epsilon_{2k}}d\Delta_{kn}.$$

Again, limiting arguments establish which terms contribute in the limit and the claim is that after summing over k and l only three terms have magnitude greater than $O_p(1)$, the term coming from the correction (A·13)

$$\frac{E(U_{1k1}V_{2k1}^*)}{E(V_{1k1}^2)},$$

the term

$$\int U_{1kj}^{\lambda_k}\lambda_k^{\epsilon_{2k}}d\Delta_{kn} \tag{A·15}$$

of which there are two, and the term

$$\int U_{1kj}^{\lambda_k\lambda_k}\lambda_k^{\epsilon_{2k}}\lambda_k^{\epsilon_{2k}}dF_k(\epsilon_{3k}). \tag{A·16}$$

The sum of (A·15) and (A·16) can be expressed as

$$2\sum_k\left[\left(\sum_j\frac{1}{\sqrt{n}}\frac{V_{1kj}}{E(V_{1k1}^2)}\right)\left\{\sum_j\frac{1}{\sqrt{n}}\left(U_{1kj}^{\lambda_k} - \frac{E(U_{1k1}^{\lambda_k}V_{1k1})}{2E(V_{1k1}^2)}V_{1kj}\right)\right\}\right]. \tag{A·17}$$

Both of the terms in parentheses have limiting distributions by application of the Central Limit theorem and their product is a quadratic form in normal random variables with mean

$$-\frac{E(U_{1k1}V_{2k1}^*)}{E(V_{1k1}^2)}.$$

This fact is best seen through recognizing the orthogonality structure in (A·17) and appealing to (1). Now in the case of estimation using U_1 then (A·17) is precisely the C term in (A·12). Consequently the asymptotic bias in the maximum likelihood estimator is the limit of

$$-\frac{1}{2}\sqrt{\overline{I}_p}(C+D)/B,$$

that is

$$-\tfrac{1}{2}\sqrt{\overline{I}}\lim_{p\to\infty}\sum_{k=1}^{p}\frac{E(U_{1k1}V_{2k1}^{*})/E(V_{1k1}^{2})}{p\overline{I}_p}.$$

However, if U_2 is used then (A·17) needs to be augmented by the correction term (A·13) and it is the fact that (A·13) exactly cancels the mean of (A·17) which verifies that C in (A·12) is mean zero when U_2 is used for estimation and hence the asymptotic bias of the root of the U_2 estimating equation is 0.

References

Andersen, E. B. (1970a). Asymptotic properties of conditional maximum likelihood estimators. *J. R. Statist. Soc. B* **32**, 283–301.

Andersen, E. B. (1970b). On Fisher's lower bound to asymptotic variances in case of infinitely many nuisance parameters. *Skand. Aktuar. J.* **53**, 78–85.

Billingsley, P. (1986). *Probability and Measure*, Second ed. New York: Wiley.

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

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5**, 815–841.

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

NEYMAN, J. & SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

PFANZAGL, J. (1993). Incidental versus random nuisance parameters. *Ann. Statist.* **21**, 1663–1691.

PORTNOY, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. consistency. *Ann. Statist.* **12**, 1298–1309.

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356–366.

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

TAK, K. M. (1982). Estimation in the presence of incidental parameters. *Can. J. Statist.* **10**, 121–132.

V. MISES, R. (1947). On the asymptotic distribution of differentiable statistical functionals. *Ann. Mat. Statist.*, 309,348.

WATERMAN, R. P. & LINDSAY, B. G. (1996a). The accuracy of projected score methods in approximating conditional scores. *Biometrika* **83**, 1,14.

Waterman, R. P. & Lindsay, B. G. (1996b). A simple and accurate method for approximate conditional inference in generalized linear models. *J. R. Statist. Soc. B* **58**, 177,188.

Table 1: *Exact 95% coverage rates for the canonical Neyman-Scott problem of Example 1 with $\lambda_i = 0$ for all i and $\sigma^2 = 1$. Columns correspond to three pivotals described in the text.*

Table 2: *Estimated coverage rates from 1000 replicates of 90% confidence intervals for ψ in the complementary log-log link model.*

Table 1

Stratum size $n = p$	1	2	3
2	0.9915	0.9770	0.9420
3	0.9866	0.9737	0.9511
4	0.9533	0.9730	0.9547
5	0.9355	0.9716	0.9538
10	0.9089	0.9620	0.9510
15	0.9021	0.9580	0.9504
20	0.8991	0.9559	0.9502

Table 2

Stratum size $n = p$	$\hat{\psi}$	$\tilde{\psi}$
10	0.840	0.930
11	0.813	0.892
12	0.831	0.897
13	0.839	0.901
14	0.835	0.894
15	0.844	0.896
16	0.838	0.898
17	0.851	0.907
18	0.823	0.883
19	0.841	0.898
20	0.830	0.893