

# Efficiency of Projected Score Methods in Rectangular Array Asymptotics

Haihong Li

Frontier Science and Technology Research Foundation, Inc.  
1244 Boylston Street, Suite 303, Chestnut Hill, MA, USA

Bruce G. Lindsay

Department of Statistics, Pennsylvania State University  
University Park, PA 16802, USA

Richard P. Waterman

Department of Statistics, University of Pennsylvania  
Philadelphia, PA 19104, USA

July 5, 2002

**Summary.** This paper considers a rectangular array asymptotic embedding for multi-stratum data sets, in which both the number of strata and the number of within-stratum replications increase, and at the same rate. It is shown that under this embedding the MLE is consistent but not efficient due to a non-zero mean in its asymptotic normal distribution. By using a projection operator on the score function, an adjusted maximum likelihood estimator can be obtained that is asymptotically unbiased and has variance that attains the Cramér-Rao lower bound. The adjusted maximum likelihood estimator can be viewed as an approximation to the conditional maximum likelihood estimator.

*Keywords:* Asymptotic theory; Neyman-Scott problem; Nuisance parameters; Projected score; 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 and Scott, 1948), demonstrate the failure of maximum likelihood estimation, and then indicate how it can be repaired through simple adjustment of 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:

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

Row  $i$  comprises a vector of independent and identically distributed observations from common density  $f(x_{ij}; \psi, \lambda_i)$ , where  $\lambda_i$  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  $\psi$  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. 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  $p/n \rightarrow c$ , 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

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

where  $i^{\psi\psi}$  is the inverse of the average (over the strata) per observation information matrix for  $\psi$ . The maximum likelihood estimator  $\hat{\psi}$  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 and Scott (1948) show that if  $n$  is held fixed then the maximum likelihood estimator need not be consistent as  $p$  tends to infinity, and even

when  $\hat{\psi}$  is consistent, it may still fail to be efficient. We use asymptotics to find good approximations for real data situations in which  $(n, p)$  are fixed. We might characterize the possibilities as follows:

1. The classical framework is the most optimistic approximation, but also the simplest.
2. The Neyman-Scott is the most pessimistic and, as a result, offers little in the way of simple results with wide applicability.
3. The rectangular array framework is intermediate in optimism. As we will show, the failings of maximum likelihood are easily corrected in this setting, leaving results nearly as simple as the classical ones.

This paper focuses on the consequences of using likelihood methods in the rectangular array setting. This is a valuable framework to consider because we will show that with this level of nuisance parameters, standard likelihood theory fails due to a potentially non zero bias term in the mean of the limiting normal distribution of the maximum likelihood estimate. However, the asymptotic variance attains the Cramér-Rao lower bound. It will be shown that a simple correction to the likelihood score equations results in an estimator which has a mean zero asymptotic normal distribution with its variance reaching the Cramér-Rao lower bound. That is, the Cramér-Rao lower bound is still a tight bound in this setting 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 under 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,b), 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. Mak (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 citations therein) concerns estimation for regression problems where the number of regression coefficients becomes large. Boos and Brownie (1995) show that in ANOVA models where the number of treatments  $p$  goes to infinity and the number of replications per treatment  $n$  remains fixed, the F test and rank statistic analogs are still meaningful in that the test statistics are asymptotically normal and distribution free as  $p \rightarrow \infty$ .

The rest of the paper is organized as follows. In Section 2 we give an example to illustrate the problem of the maximum likelihood estimator in the rectangular array setting. Section 3 describes procedures to obtain projected scores. In Section 4 we consider the asymptotic properties of projected scores in one stratum. The asymptotic results of estimators based on the second order projected scores are presented in Section 5. Two numerical examples are considered in Section 6, and Section 7 contains discussion. The appendix contains the regularity conditions and proof for Theorem 5.2.

## 2 An example

We consider the problem of estimating the variance  $\sigma^2$  in one-way analysis of variance (ANOVA).

The random variable  $X_{ij}$  follows a normal distribution with mean  $\mu_i$  and common variance  $\sigma^2$ ,  $i = 1, \dots, p$ ,  $j = 1, \dots, n$ . Regardless of the asymptotic embedding, the maximum likelihood estimator of  $\sigma^2$  is  $\hat{\sigma}^2 = (np)^{-1} \sum_{ij} (x_{ij} - \bar{x}_i)^2$ , but the limit in probability of this estimator depends on the asymptotic formulation. In both the classical and rectangular array settings the maximum likelihood estimator is consistent. However, in the Neyman-Scott setting the limiting value of  $\hat{\sigma}^2$  is  $(n-1)\sigma^2/n$ , so the maximum likelihood estimator is inconsistent. For small  $n$ , for example when  $n = 2$ , this is potentially disastrous if ignored.

In the classical asymptotics  $\hat{\sigma}^2$  is efficient as well as consistent, by which is meant  $\sqrt{np}(\hat{\sigma}^2 - \sigma^2) \rightarrow N(0, i^{\sigma^2\sigma^2})$ , as  $n \rightarrow \infty$ , where  $i^{\sigma^2\sigma^2} = 2\sigma^4$  is the inverse of the Fisher information about  $\sigma^2$  in the presence of  $\mu$ , calculated as  $(i_{\sigma^2\sigma^2} - i_{\sigma^2\mu}i_{\mu\mu}^{-1}i_{\mu\sigma^2})^{-1}$  from the elements of the Fisher information matrix. In Neyman-Scott asymptotics the maximum likelihood estimator fails to be consistent so 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  $(n, p) \rightarrow \infty$ , 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. We will generalize this result to show that in a rectangular array setting, maximum likelihood has a bias term in the mean, but the asymptotic variance is obtainable from the Fisher information matrix.

## 3 Notation and tools

We assume the following distributional setting: let  $\Psi$  and  $\Lambda$  be two subsets of real numbers  $\mathbb{R}^1$ , with  $\psi \in \Psi$  and  $\lambda_i \in \Lambda$ . Let  $\mathcal{X}$  be the common sample space for  $X_{ij}$ ,  $i = 1, \dots, p$ ,  $j =$

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

This section will focus on a single stratum of the rectangular array. In Section 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 stratum there is only one nuisance parameter, the index  $i$  is fixed and is therefore omitted.

The technique we will use to repair the maximum likelihood estimating equations is the projected score approach (Small and McLeish, 1989; Waterman and Lindsay, 1996a). The aim of the procedure is to reduce the impact of the nuisance parameters on the estimating equation for  $\psi$ . This is achieved by subtracting from the  $\psi$ -score function  $U_0$  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  $k = 1, \dots, t$ , and  $f^{(k)}$  is the  $k$ -th order derivative of  $f$  with respect to the nuisance parameter. It was shown that under regularity conditions  $\mathcal{V}_t$  approaches the ‘‘E-sufficient’’ subspace for the nuisance parameter as  $t \rightarrow \infty$  (Small and McLeish, 1994). The basis formed by  $V_k, k = 1, \dots$ , is referred to as the Bhattacharyya basis.

Denoting the projection operator onto  $\mathcal{V}_t$  by  $\Pi_t$ , the projected scores are defined as  $U_t = U_0 - \Pi_t U_0$ . If a complete and sufficient statistic,  $T$ , exists for  $\lambda$  then the space spanned by  $\mathcal{V}_t$  as  $t \rightarrow \infty$  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).

In this paper we consider projected scores of up to second order. Let  $V_1^*, \dots, V_t^*$  be an orthogonalized version of the Bhattacharyya basis obtained using the Gram-Schmidt algorithm, that is,

$$\begin{aligned} V_1^* &= V_1, \\ V_2^* &= V_2 - \text{projection of } V_2 \text{ on } V_1^* = V_2 - \frac{E(V_2 V_1)}{E(V_1^2)} V_1, \\ &\dots \\ V_t^* &= V_t - \text{projection of } V_t \text{ on } V_1^*, V_2^*, \dots, V_{t-1}^*. \end{aligned}$$

Using this orthogonalized basis, the first and second order projected scores can be written

as

$$\begin{aligned}
 U_1 &= U_0 - \rho_1 V_1^* = U_0 - \frac{E(U_0 V_1)}{E(V_1^2)} V_1, \\
 U_2 &= U_0 - \rho_1 V_1^* - \rho_2 V_2^* = U_1 - \frac{E(U_1 V_2^*)}{E(V_2^*)} V_2^*.
 \end{aligned}$$

We use  $U_{1ij}$  to denote  $U_1$  for the  $j$ -th observation in the  $i$ -th stratum.  $U_{2ij}$  and  $V_{tij}$  are defined in a similar fashion.

Building from these within-stratum projected score functions, this paper considers the derived  $U_1$  and  $U_2$  projected score functions for the rectangular array. The maximum likelihood estimate  $\hat{\psi}_1$  and adjusted maximum likelihood estimate  $\hat{\psi}_2$  are obtained by solving estimating equations

$$U_1(\psi, \hat{\lambda}_1(\psi), \dots, \hat{\lambda}_p(\psi)) = 0$$

and

$$U_2(\psi, \hat{\lambda}_1(\psi), \dots, \hat{\lambda}_p(\psi)) = 0$$

respectively, where  $\hat{\lambda}_i(\psi)$  is the maximum likelihood estimator of  $\lambda_i$  at fixed  $\psi$ . The main aim here is to show that  $\hat{\psi}_2$  is a fully efficient estimate in the rectangular array. We note that the use of  $U_2$  is just one method of removing bias from maximum likelihood estimators. Other bias-reducing methods, for example modified profile likelihood (Barndorff-Nielsen, 1980, 1983), adjusted scores (McCullagh and Tibshirani, 1990), and integrated likelihoods (Kalbfleisch and Sprott, 1970; Berger et al., 1999; Reid, 1996), should have the same efficiency property.

## 4 Limiting distributions in one stratum

The estimation of the nuisance parameters has a cost. In this section we investigate the cost in bias and information loss based on using second order asymptotics as  $n \rightarrow \infty$  within a single stratum  $i$ , for  $i = 1, \dots, p$ . The asymptotic results are obtained using standard Taylor series arguments. The results have multivariate analogs, but only univariate  $\psi$  will be considered here.

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 function. The proof can be found in Waterman (1993).

Under regularity conditions, we have

$$U_1(\psi, \hat{\lambda}_\psi) - U_1(\psi, \lambda) \xrightarrow{d} Y_1 + \frac{1}{2}\kappa, \quad (1)$$

$$U_2(\psi, \hat{\lambda}_\psi) - U_1(\psi, \lambda) \xrightarrow{d} Y_1, \quad (2)$$

$$U_2(\psi, \hat{\lambda}_\psi) - U_2(\psi, \lambda) \xrightarrow{d} Y_2. \quad (3)$$

Here both  $Y_1$  and  $Y_2$  are mean zero quadratic forms of bivariate normal random vectors with  $\text{Var}(Y_1) > \text{Var}(Y_2)$ ;  $\kappa$  equals  $-E(U_{1i1}V_{2i1})/E(V_{1i1}^2)$ , and has an interpretation as a model curvature. The equations have the following interpretations: (1) shows the inherent bias of  $\frac{1}{2}\kappa$  in the estimated  $U_1$  equation, which leads to the asymptotic bias of the maximum likelihood estimator in the rectangular array setting. Equation (2) 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 (3) 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 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, while the order of  $\hat{U}_2 - U_2$  is small enough so that the bias is small even after being accumulated over strata. For the square array setting, we are essentially looking for a “second order locally E-ancillary” estimating equation, using the terminology of Small and McLeish (1988).

## 5 Main theorem

In Theorem 5.2 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-stratum projected scores. The function  $U_{ti}$  is the  $t$ -th order projected score from the  $i$ -th stratum.

**Lemma 5.1.** *In an independent stratum sampling scheme, both the projective scores and the information are additive over the strata:*

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

*Proof.* The first result is obvious. For the second result we only need to notice the information unbiasedness property of  $U_t$ , which says that  $E(U_t^\psi) + E(U_t^2) = 0$ , where  $U_t^\psi$  is the partial derivative  $\frac{\partial}{\partial \psi} U_t$ . The information unbiasedness was shown in Waterman and Lindsay (1996a).  $\square$

Consider the marginal score function  $U_1$  and second order projected score function  $U_2 = U_1 - \rho_2 V_2$ . In both functions we obtain the estimate  $\lambda_\psi$  of the unknown nuisance parameter  $\lambda$  at fixed  $\psi$  values by solving  $V_1(\psi, \lambda) = 0$ . Let  $\hat{\psi}_1$  be the solution of  $U_1(\psi, \lambda_\psi) = 0$ , and let  $\hat{\psi}_2$  be the solution of  $U_2(\psi, \lambda_\psi) = 0$ .

We shall show that, under regularity conditions, the maximum likelihood estimator  $\hat{\psi}_1$  is inefficient, while the adjusted maximum likelihood estimator  $\hat{\psi}_2$  is efficient, although both are consistent.

The full list of regularity conditions is long and is given in Appendix A. One salient feature is that since we are putting the problem in an asymptotic setting, we need to use future unobserved values of  $\lambda_i$ . For a good approximation to the fixed sample size case, the future sequence  $\{\lambda_i\}$  should be similar to the existing values rather than behaving unrealistically and pathologically. In many cases it is reasonable to model  $\{\lambda_i\}$  by assuming it has the stochastic properties of a sample from some distribution.

As a simple counterexample, we consider the setup of the earlier one-way ANOVA but here we estimate the common mean  $\mu$  with different variances  $\sigma_i^2$  as nuisance parameters. For a balanced design the exact maximum likelihood estimator is  $\hat{\mu} = (\sum_{i=1}^p w_i \bar{X}_i) / (\sum_{i=1}^p w_i)$ , where  $w_i = 1/\sigma_i^2$ . The variance of  $\hat{\mu}$  is  $1/(np\bar{w})$ . It's easy to see that if the future values of  $\sigma_i^2$  are very different from the observed ones, the asymptotic variance will be a bad approximation to the finite sample variance.

The behavior of the sequence  $\{\lambda_i\}$  should be characterized as part of the asymptotic scheme to ensure some uniformity. The conditions we give may seem strong, but the whole setup in the square array is more general (and more pessimistic) than the classical asymptotic embedding that fixes the number of strata.

**Theorem 5.2.** *Let  $\hat{\psi}_1$  and  $\hat{\psi}_2$  denote the MLE and the adjusted MLE respectively. Under the assumptions stated in Appendix A, we have*

$$\sqrt{np}(\hat{\psi}_1 - \psi) \xrightarrow{d} N(b, \bar{I}^{-1}),$$

and

$$\sqrt{np}(\hat{\psi}_2 - \psi) \xrightarrow{d} N(0, \bar{I}^{-1}) \quad \text{as } p = n \rightarrow \infty,$$

where

$$\bar{I} = \lim_{p \rightarrow \infty} \bar{I}_p = \lim_{p \rightarrow \infty} \frac{\sum_{i=1}^p E(U_{1ij}^2)}{p},$$

and

$$b = -\frac{1}{2} \lim_{p \rightarrow \infty} \sum_{k=1}^p \frac{E(U_{1kj} V_{2kj}) / E(V_{1kj}^2)}{p \bar{I}_p}. \quad (4)$$

*Proof.* Let  $F = (F_1, \dots, F_p)$  be the vector of distribution functions, and  $\hat{F} = (\hat{F}_1, \dots, \hat{F}_p)$ , where  $\hat{F}_i$  is the empirical distribution function for stratum  $i$ . Define

$$F(\epsilon) = F + \epsilon \sqrt{n} (\hat{F} - F), \quad \text{for } \epsilon \in [0, n^{-1/2}].$$

We can see that  $F(0) = F$ ,  $F(n^{-1/2}) = \hat{F}$ , and any  $F(\epsilon)$  is a point on the line segment from  $F$  to  $\hat{F}$ .

For each fixed  $\psi$  and  $\epsilon$ , let  $\lambda_i\{\psi, F_i(\epsilon)\}$  be the solution to the estimating function

$$V_{1i}(\epsilon) = \int V_{1i1}[\psi, \lambda_i\{\psi, F_i(\epsilon)\}] dF_i(\epsilon) = 0.$$

For each  $\epsilon \in [0, n^{-1/2}]$ , we define the functional  $\psi_1[F(\epsilon)]$  to be the solution for  $\psi$  in

$$U_1(\epsilon) = \sum_{i=1}^p U_{1i}(\epsilon) = \sum_{i=1}^p \int U_{1i1}[\psi(F(\epsilon)), \lambda_i\{\psi(F(\epsilon)), F_i(\epsilon)\}] dF_i(\epsilon) = 0,$$

and define the functional  $\psi_2[F(\epsilon)]$  to be the solution of

$$U_2(\epsilon) = \sum_{i=1}^p U_{2i}(\epsilon) = 0.$$

The functional representation of  $\sum_{i=1}^p U_{2i}(\epsilon)$  will be derived in the appendix when we give a detailed proof.

We can see immediately that

$$\psi_1[F(0)] = \psi_2[F(0)] = \psi, \quad \psi_1[F(n^{-1/2})] = \hat{\psi}_1, \quad \psi_2[F(n^{-1/2})] = \hat{\psi}_2.$$

For both  $\psi_1$  and  $\psi_2$ , we use a Taylor expansion in  $\epsilon$  of  $\psi(\hat{F})$  at  $F_0$ ,

$$\psi_l(\hat{F}) - \psi_l(F_0) = n^{-\frac{1}{2}} \left. \frac{d\psi_l[F(\epsilon)]}{d\epsilon} \right|_{\epsilon=0} + \frac{1}{2} n^{-1} \left. \frac{d^2\psi_l[F(\epsilon)]}{d\epsilon^2} \right|_{\epsilon=0} + \frac{1}{6} n^{-\frac{3}{2}} \left. \frac{d^3\psi_l[F(\epsilon)]}{d\epsilon^3} \right|_{\epsilon=\epsilon^*}, \quad l = 1, 2. \quad (5)$$

where  $\epsilon^*$  is between 0 and  $n^{-1/2}$ .

The theorem will be proved if we verify the following statements, as we will do in the appendix.

1. The first derivative term is:

$$\frac{d\psi_1[F(\epsilon)]}{d\epsilon} = \frac{d\psi_2[F(\epsilon)]}{d\epsilon} = -\frac{n^{-1/2} \sum_{i=1}^p \sum_{j=1}^n U_{1ij}}{\sum_{i=1}^p E(U_{1i1}^\psi)}.$$

Therefore

$$\sqrt{np} \sum_{i=1}^p \frac{1}{\sqrt{n}} \frac{d\psi_l[F(\epsilon)]}{d\epsilon} \Big|_{\epsilon=0} = \frac{1}{\sqrt{pn}} \frac{\sum_{i=1}^p \sum_{j=1}^n U_{1ij}}{\frac{1}{p} \sum_{i=1}^p E(U_{1i1}^2)} \xrightarrow{d} N(0, \bar{I}^{-1}), \quad l = 1, 2.$$

2. The second order derivatives satisfy:

$$\frac{1}{2} \frac{d^2\psi_1[F(\epsilon)]}{d\epsilon^2} \Big|_{\epsilon=0} \xrightarrow{p} b, \quad \frac{1}{2} \frac{d^2\psi_2[F(\epsilon)]}{d\epsilon^2} \Big|_{\epsilon=0} \xrightarrow{p} 0.$$

3. The remainder terms for both  $\psi_1(\cdot)$  and  $\psi_2(\cdot)$  in equation (5) are small enough to be ignored. We will show that  $\sup_{\epsilon \in [0, n^{-1/2}]} \left| \frac{d^3\psi_l[F(\epsilon)]}{d\epsilon^3} \right| = O_p(1)$  for  $l = 1, 2$ .

We complete the proof by combining the above results. □

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 stratum. The elements of  $b$  can be extracted from the within stratum *extended information matrix* for a single observation, the matrix being defined as  $M_2 = E\{(U_{0i1}, V_{1i1}, V_{2i1})^T (U_{0i1}, V_{1i1}, V_{2i1})\}$ . Here

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

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

The above example is somewhat special because the nuisance parameters  $\lambda_i$  do not appear in the information and hence they are free to take any value. In more general cases we require constraints on the sequence of nuisance parameters to ensure regularity.

## 6 Numerical results

Here we present the results of some simulations that indicate the effect of bias in the limiting distribution of the maximum likelihood estimator on confidence intervals for the parameter of interest.

We start with the standard Neyman-Scott problem. Exact coverage rates were calculated for confidence intervals based on 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 (Waterman and Lindsay, 1996b), and the third is for an information corrected version of  $U_2$ . See Lindsay and Waterman (1999) for methods for correcting the information in  $U_2$ . Exact coverage rates for nominal 95% equi-tailed confidence intervals were calculated using the numerical integration function `integrate` in R, and are reported in Table 1. For convenience all the nuisance parameters were set to  $\lambda_i = 0$  and  $\sigma^2 = 1$ . The asymptotic pivotals are

$$Z_1 = \frac{\sqrt{n p}(\hat{\psi} - \psi)}{\sqrt{2}\psi} \quad Z_2 = \frac{\sqrt{n p}(\tilde{\psi} - \psi)}{\sqrt{2}\psi} \quad Z_3 = a(\kappa, i) \frac{\sqrt{n p}(\tilde{\psi} - \psi)}{\sqrt{2}\psi}.$$

where  $a(\kappa, i) = 1 - 1/2n$ .

As both  $n$  and  $p$  increase, the maximum likelihood estimator coverage rates, the first number of each cell, are systematically biased, whereas the  $U_2$  coverage rates, the second number, approach the correct nominal value of 0.95. However, the extremely accurate properties of the information-corrected estimator warrant further investigation.

As a second numerical example we present an extension of the binary matched pairs experiment. Let  $X_i$  and  $Y_i, i = 1, \dots, p$  be independent binomial random variables based on  $n = p$  trials and with success probabilities  $q(0, \lambda_i)$  and  $q(\psi, \lambda_i)$  respectively. We take

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

This 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  $\lambda_i$  do not apply. The simulation, carried out in FORTRAN, consisted of performing 1000 replicates

Table 1: Exact coverage rates of 95% confidence intervals based on the three pivots for the standard Neyman-Scott problem

<i>Stratum size</i>	<i>n</i>	2	3	4	5	10	15	20
<i>Number of strata p</i>								
2	$Z_1$	99.2	98.8	98.5	98.0	96.4	95.9	95.7
	$Z_2$	90.8	92.6	93.5	93.9	94.5	94.7	94.7
	$Z_3$	94.2	94.9	95.2	95.5	95.5	95.4	95.3
3	$Z_1$	99.5	98.7	97.3	96.7	95.6	95.4	95.3
	$Z_2$	90.6	92.5	93.2	93.5	94.3	94.5	94.6
	$Z_3$	94.3	95.1	95.5	95.5	95.4	95.3	95.2
4	$Z_1$	99.4	96.1	95.3	95.1	94.9	94.9	94.9
	$Z_2$	90.5	91.9	92.7	93.2	94.1	94.4	94.6
	$Z_3$	94.4	95.3	95.5	95.5	95.3	95.2	95.2
5	$Z_1$	93.9	93.1	93.3	93.6	94.2	94.5	94.6
	$Z_2$	89.3	91.4	92.4	93.0	94.1	94.4	94.5
	$Z_3$	94.5	95.3	95.4	95.4	95.2	95.2	95.1
10	$Z_1$	66.7	78.6	83.5	86.1	90.9	92.3	93.0
	$Z_2$	86.2	90.2	91.8	92.5	93.9	94.3	94.5
	$Z_3$	94.7	95.1	95.2	95.2	95.1	95.1	95.1
15	$Z_1$	46.5	65.4	74.0	78.9	87.6	90.2	91.5
	$Z_2$	85.2	89.8	91.5	92.4	93.8	94.2	94.4
	$Z_3$	94.5	94.9	95.0	95.1	95.1	95.1	95.0
20	$Z_1$	31.6	53.6	65.1	71.8	84.3	88.1	89.9
	$Z_2$	84.7	89.6	91.4	92.3	93.8	94.2	94.4
	$Z_3$	94.3	94.8	95.0	95.0	95.0	95.0	95.0

Table 2: Point estimates of  $\hat{\psi}$  and  $\tilde{\psi}$  with standard errors for the paired binomial problem:  $\hat{\psi}$  is the upper value and  $\tilde{\psi}$  is the lower value in each cell; the true parameter value is 0.75

		$n = p$		$n = p/2$		$n = 2p$	
$n$	$p$	$\hat{\psi}$ and $\tilde{\psi}$		$n$	$p$	$\hat{\psi}$ and $\tilde{\psi}$	
10	10	0.824(0.213)	0.765(0.194)	10	20	0.826(0.146)	0.768(0.133)
12	12	0.810(0.176)	0.763(0.164)	11	22	0.824(0.138)	0.771(0.127)
14	14	0.788(0.143)	0.749(0.134)	12	24	0.806(0.118)	0.759(0.110)
16	16	0.787(0.122)	0.754(0.116)	13	26	0.788(0.106)	0.747(0.099)
18	18	0.779(0.112)	0.750(0.107)	14	28	0.801(0.100)	0.762(0.094)
20	20	0.778(0.099)	0.752(0.095)	15	30	0.782(0.094)	0.747(0.088)
				20	10	0.782(0.139)	0.755(0.134)
				22	11	0.774(0.125)	0.750(0.121)
				24	12	0.778(0.116)	0.756(0.112)
				26	13	0.771(0.107)	0.751(0.104)
				28	14	0.775(0.098)	0.757(0.096)
				30	15	0.772(0.087)	0.755(0.085)

at each level of  $(n, p)$ . The nuisance parameters were set equal at 0.0, while the parameter of interest was set to 0.75. Both the maximum likelihood estimator  $\hat{\psi}$  and the adjusted maximum likelihood estimator  $\tilde{\psi}$  were calculated. Their means and standard errors (in parentheses) are reported in Table 2.

The table again corroborates the assertion of the main theorem. The maximum likelihood estimator performs badly, with bigger bias and bigger standard error. But the corrected maximum likelihood estimator consistently performs better. The biases are large in the  $n = p/2$  case, and small in the  $n = 2p$  case. This is not surprising, since the former case is closer to the Neyman-Scott setup while the latter case is closer to the classical setup. Theorem 5.2 can be extended to the case  $n = cp$ , and the asymptotic bias can be computed and compared with the simulation results here. The bias of the MLE is a function of the curvature  $\kappa$ , 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.

## 7 Discussion

In the paper we have only considered balanced within-stratum sample sizes. The results still hold in the case of unbalanced stratum sample sizes so long as no stratum 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-stratum sample size grows at a different rate than the number of strata, that is  $p = n^\alpha$ , for example if  $\alpha = 2$  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, and that the asymptotic variance will exceed the Cramér-Rao lower bound.

This general setting of within-stratum sample size growing at different rate than the number of strata was studied by Portnoy (1984, 1988), under different model assumptions: there is no partition into parameters of interest and nuisance parameters; data are samples from distributions that depend on all the parameters; and inference is needed for all the parameters. He obtained moment conditions for consistency and asymptotic normality of the parameter vector, which is increasing in dimension itself. Our model is a generalization in the sense that the data are not coming from identical distributions, each observation is a sample from a distribution that depends on only a finite subset of the parameters.

The paper concentrates on univariate  $\psi$  and  $\lambda_i$  for convenience of exposition. One of the strengths of the methodology is its straightforward extension to the multivariate case. Waterman and Lindsay (1996b) gave an algorithm for solving the second-order projected score estimator for generalized linear models with canonical link functions and known dispersion parameters.

Finally, we note that an alternative to modeling the  $\lambda_i$  as nuisance parameters is to model them as a random sample of latent variables from some distribution, itself modeled using a parametric or nonparametric formulation (Lindsay, 1995). This setting provides yet a fourth asymptotic embedding for our problem. A detailed comparison of this embedding with the others is beyond the scope of this paper, but would offer further insights into nuisance parameter problem.

## Acknowledgments

We are very grateful to the joint Editor and the referee for helpful comments and suggestions which improved on our manuscript substantially. Lindsay's research was partially supported by National Science Foundation grant DMS 9870193.

## A REGULARITY CONDITIONS

For Theorem 5.2, we make the following assumptions:

1. We assume that both  $\Psi$  and  $\Lambda_i$  are open intervals in  $\mathbb{R}^1$ , with  $\psi \in \Psi$  and  $\lambda_i \in \Lambda_i$ .
2. For the smoothness of the likelihood, we assume the second order derivatives  $U_{1ij}^{\psi\psi}$ ,  $U_{1ij}^{\psi\lambda_i}$ ,  $V_{1ij}^{\psi\lambda_i}$ ,  $V_{1ij}^{\lambda_i\lambda_i}$  exist for all  $i, j$  and all  $x$ .
3. For each  $\psi_0 \in \Psi$  and  $\lambda_{i0} \in \Lambda_i$ , there exist functions  $g_i(x)$ ,  $h_i(x)$ , and  $H(x)$  (possibly depending on  $(\psi, \lambda_{i0})$ ) for each stratum  $i$ , such that for  $\psi$  and  $\lambda_i$  in their neighborhood  $N(\psi_0, \lambda_{i0})$ , the relations

$$\left| \frac{\partial f_i(x; \psi, \lambda_i)}{\partial \psi} \right| \leq g_i(x), \quad \left| \frac{\partial f_i(x; \psi, \lambda_i)}{\partial \lambda_i} \right| \leq g_i(x),$$

$$\left| \frac{\partial^2 f_i(x; \psi, \lambda_i)}{\partial \psi^2} \right|, \left| \frac{\partial^2 f_i(x; \psi, \lambda_i)}{\partial \psi \partial \lambda_i} \right|, \left| \frac{\partial^2 f_i(x; \psi, \lambda_i)}{\partial \lambda_i^2} \right| \leq h_i(x),$$

and

$$\left| U_{1ij}^{\psi\psi} \right|, \left| U_{1ij}^{\psi\lambda_i} \right|, \left| U_{1ij}^{\lambda_i\lambda_i} \right|, \left| V_{1ij}^{\psi\psi} \right|, \left| V_{1ij}^{\psi\lambda_i} \right|, \left| V_{1ij}^{\lambda_i\lambda_i} \right| \leq H(x)$$

hold, for all  $x$ , and

$$\int g_i(x) dx < \infty, \quad \int h_i(x) dx < \infty, \quad E_{(\psi, \lambda_i)}\{H(x)\} < \infty \quad \text{for } (\psi, \lambda_i) \in N(\psi_0, \lambda_{i0}).$$

We assume the upper bound  $H$  on the second order derivatives to be independent of stratum.

4. The limits

$$\bar{I} = \lim_{p \rightarrow \infty} \bar{I}_p = \lim_{p \rightarrow \infty} \frac{\sum_{i=1}^p E(U_{1ij}^2)}{p} > 0, \quad b = -\frac{1}{2} \lim_{p \rightarrow \infty} \sum_{k=1}^p \frac{E(U_{1kj}V_{2kj})/E(V_{1kj}^2)}{p\bar{I}_p},$$

and

$$\lim_{p \rightarrow \infty} \frac{\sum_{i=1}^p [E(V_{1ij}^2)]^{-1}}{p} > 0$$

exist for all  $\psi_0 \in \Psi$  and  $\lambda_i \in \Lambda_i$ . Here we have implicitly assumed that the expectations involved are finite. The third condition is to ensure that the information on the nuisance parameter will not become too small as  $p$  increases.

5. A Lindeberg condition on  $U_1, V_1$ , and  $U_1^\lambda$  is assumed to ensure that the central limit theorem can be used, namely,

$$\frac{1}{p\bar{I}_p} \sum_{i=1}^p \int_{\left| \frac{\sum_j U_{1ij}}{\sqrt{pn\bar{I}_p}} \right| \geq \epsilon} \left( \frac{\sum_j U_{1ij}}{\sqrt{n}} \right)^2 dF_i \rightarrow 0, \quad \text{where } \bar{I}_p = \frac{1}{p} \sum_{i=1}^p E(U_{1i1}^2),$$

$$\frac{1}{p\bar{J}_p} \sum_{i=1}^p \int_{\left| \frac{\sum_j V_{1ij}}{\sqrt{pn\bar{J}_p}} \right| \geq \epsilon} \left( \frac{\sum_j V_{1ij}}{\sqrt{n}} \right)^2 dF_i \rightarrow 0, \quad \text{where } \bar{J}_p = \frac{1}{p} \sum_{i=1}^p E(V_{1i1}^2),$$

$$\frac{1}{p\bar{I}_p^\lambda} \sum_{i=1}^p \int_{\left| \frac{\sum_j U_{1ij}^{\lambda_i}}{\sqrt{pn\bar{I}_p^\lambda}} \right| \geq \epsilon} \left( \frac{\sum_j U_{1ij}^{\lambda_i}}{\sqrt{n}} \right)^2 dF_i \rightarrow 0, \quad \text{where } \bar{I}_p^\lambda = \frac{1}{p} \sum_{i=1}^p E(U_{1i1}^{\lambda_i^2}).$$

6. For each stratum  $i$ , there exists a function  $M_i(x)$ , possibly depending on  $(\psi_0, \lambda_{i0})$ , such that  $|U_{1ij}^{\psi\psi}(x; \psi, \lambda_i)| \leq M_i(x)$  and  $|U_{1ij}^{\psi\lambda_i}(x; \psi, \lambda_i)| \leq M_i(x)$ .
7. The  $M_i(x)$  are integrable:  $E_{(\psi, \lambda_i)} M_i(x) < \infty$ .
8. There exists a real number  $M$  such that  $\frac{1}{p} \sum_{i=1}^p [E_{(\psi, \lambda_i)} M_i(x)]^2 \leq M$  as long as  $p$  is sufficiently large. This implies that  $\frac{1}{p} \sum_{i=1}^p E_{(\psi, \lambda_i)} M_i(x)$  is also bounded.
9. Let  $\hat{\lambda}_i^\psi$  be the maximum likelihood estimator of  $\lambda_i$  for fixed  $\psi$ . It will be assumed that the mean square of the sequence  $\hat{\lambda}_i^\psi - \lambda_i$  converges to zero in probability:

$$\frac{1}{p} \sum_{i=1}^p (\hat{\lambda}_i^\psi - \lambda_i)^2 \xrightarrow{p} 0.$$

If we treat the sequence  $\{\lambda_i\}$  as following a certain distribution  $Q$ , the resulting  $X$  are a random sample from the mixture distribution  $\int f(x; \psi, \lambda) dQ$ , and the above condition can be rewritten as

$$\hat{E}_Q(\hat{\Lambda} - \Lambda)^2 \xrightarrow{p} 0,$$

which means that on average the square error  $(\hat{\Lambda} - \Lambda)^2$  goes to zero.

10. We impose some conditions to ensure the law of large numbers valid for the two triangular arrays  $X_{pi} = \frac{1}{p} \frac{\sum_{j=1}^n M_i(x_{ij})}{n} \cdot E(M_i)$  and  $Y_{pi} = \frac{1}{p} \left( \frac{\sum_{j=1}^n M_i(x_{ij})}{n} \right)^2$ , where  $i = 1, \dots, p$  and  $p \rightarrow \infty$ . One set of conditions is given by Feller (1966, p. 316):

$$\sum_{i=1}^p P\{|X_{pi}| > \eta\} \rightarrow 0, \quad \text{and} \quad \sum_{i=1}^p \text{Var}(X'_{pi}) \rightarrow 0.$$

for each  $\eta > 0$  and each truncation level  $s > 0$ , where  $X'$  is the truncation of random variable  $X$ , that is,

$$X' = \begin{cases} X & \text{if } |X| \leq s, \\ 0 & \text{if } |X| > s. \end{cases}$$

Conditions for  $Y_{pi}$  are the same.

## B PROOF OF THE MAIN THEOREM

### B.1 Results for later use

This section provides some key facts about the maximum likelihood functional  $\lambda_i(\psi, F_i)$  that solves the MLE score equation for the nuisance parameter when the parameter of interest  $\psi$  is fixed.

In the  $i$ th stratum  $\lambda_i(\psi, F_i(\epsilon))$  solves

$$\int V_{1i1}[\psi, \lambda_i\{\psi, F_i(\epsilon)\}] dF_i(\epsilon) = 0. \quad (\text{B}\cdot 1)$$

We take derivatives of (B.1) with respect to  $\psi$  and  $\epsilon$ . The  $\psi$ -derivative is:

$$\int V_{1i1}^\psi[\psi, \lambda_i\{\psi, F_i(\epsilon)\}] dF_i(\epsilon) + \lambda_i^\psi\{\psi, F_i(\epsilon)\} \int V_{1i1}^{\lambda_i}[\psi, \lambda_i\{\psi, F_i(\epsilon)\}] dF_i(\epsilon).$$

The  $\epsilon$ -derivative is:

$$\lambda_i^\epsilon\{\psi, F_i(\epsilon)\} \int V_{1i1}^{\lambda_i}[\psi, \lambda_i\{\psi, F_i(\epsilon)\}] dF_i(\epsilon) + \int V_{1i1}[\psi, \lambda_i\{\psi, F_i(\epsilon)\}] d\Delta_{in},$$

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

Equating these equations to zero and solving for the derivatives of  $\lambda_i$  evaluated at  $\epsilon = 0$  gives

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

and

$$\lambda_i^\epsilon = -\frac{\int V_{1i1} d\Delta_{in}}{\int V_{1i1}^{\lambda_i} dF_i} = \frac{\sum_j V_{1ij}}{\sqrt{n}E(V_{1i1}^2)}. \quad (\text{B}\cdot 2)$$

Thus in future expansions,  $\lambda_i^\psi$  is  $O(1)$  but  $\lambda_i^\epsilon$  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 (B.2).

## B.2 Expressing $U_{2i}$ in functional form

In the  $i$ th 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  $\rho_i$  is the appropriate regression coefficient. By definition  $V_{2i}^* = V_{2i} - \tau_i V_{1i}$  where again  $\tau_i$  makes the appropriate orthogonality adjustment.

By definition  $\rho_i = E(U_{1i} V_{2i}^*) / E(V_{2i}^{*2})$ , which can be written as  $n^{-1} E(U_{1i1} V_{2i1}) / 2 \{E(V_{1i1}^2)\}^2 + O(n^{-2})$ . This expression can be obtained using the fact that

$$V_{2i}^* = V_{2i} - \tau_i V_{1i} = V_{1i}^{\lambda_i} + V_{1i}^2 - \tau_i V_{1i} = \sum_j V_{2ij} + \sum_{j_1 \neq j_2} V_{1ij_1} V_{1ij_2} - \frac{E(V_{2i} V_{1i})}{E(V_{1i}^2)} \sum_j V_{1ij}, \quad (\text{B.3})$$

together with the independence of observations, the orthogonality between  $U_{1i}$  and  $V_{1i}$ , and zero-unbiasedness of  $U_{1ij}$ ,  $V_{1ij}$ , and  $V_{2ij}$ .

Also from equation (B.3),  $U_{2i}$  estimating function can be written in functional form as (where  $\tilde{\rho}_i = E(U_{1i1} V_{2i1}) / 2 \{E(V_{1i1}^2)\}^2$ ),

$$\frac{1}{n} U_{2i} = \int U_{1i1} d\hat{F}_i - \left[ \frac{1}{n^2} \tilde{\rho}_i + O\left(\frac{1}{n^3}\right) \right] \left[ n \int V_{1i1}^{\lambda_i} d\hat{F}_i + n^2 \left( \int V_{1i1} d\hat{F}_i \right)^2 + \tau_i n \int V_{1i1} d\hat{F}_i \right].$$

For any  $\epsilon \in [0, n^{-1/2}]$ , we define

$$U_{2i}(\epsilon) = \int U_{1i1} dF_i(\epsilon) - \left[ \epsilon^4 \tilde{\rho}_i + \epsilon^6 W_i(\epsilon) \right] \left[ \frac{1}{\epsilon^2} \int V_{1i1}^{\lambda_i} dF_i(\epsilon) + \frac{1}{\epsilon^4} \left( \int V_{1i1} dF_i(\epsilon) \right)^2 + \tau_i \frac{1}{\epsilon^2} \int V_{1i1} dF_i(\epsilon) \right],$$

where  $W_i(\epsilon)$  is some function of  $\epsilon$  with order  $O(1)$ . Since we will evaluate functions at the maximum likelihood estimator for the nuisance parameters, the last two terms in the large bracket will be zero.

Since we are evaluating derivatives up to the 3rd order only, any term involving  $\epsilon^4$  will not appear when  $\epsilon$  is set to zero. Therefore we will use the simpler equivalent expression to evaluate the derivatives at zero:

$$\tilde{U}_{2i}(\epsilon) = \int U_{1i1} dF_i(\epsilon) - \tilde{\rho}_i \epsilon^2 \int V_{1i1}^{\lambda_i} dF_i(\epsilon).$$

We notice that  $U_{2i}^*(\epsilon)$  equals to  $U_{1i}(\epsilon)$  minus a correction term.

Combining the  $p$  strata we get the equation

$$\sum_{i=1}^p \tilde{U}_{2i}(\epsilon) = \sum_{i=1}^p \left\{ \int U_{1i1}[\psi(\epsilon), \lambda_i(\psi(\epsilon), \epsilon)] dF_i(\epsilon) - \tilde{\rho}_i \epsilon^2 \int V_{1i1}^{\lambda_i}[\psi(\epsilon), \lambda_i(\psi(\epsilon), \epsilon)] dF_i(\epsilon) \right\} = 0.$$

### B.3 First order derivatives

The first order derivative of  $U_1(\epsilon)$  is

$$\begin{aligned} \frac{\partial}{\partial \epsilon} U_1(\epsilon) &= \sum_{i=1}^p \left\{ \int \left( U_{1i1}^{\psi} \psi^\epsilon + U_{1i1}^{\lambda_i} \lambda_i^\psi \psi^\epsilon \right) dF_i(\epsilon) \right. \\ &\quad + \int U_{1i1}^{\lambda_i} \lambda_i^\epsilon dF_i(\epsilon) \\ &\quad \left. + \int U_{1i1} d\Delta_i \right\} \quad \text{where } \Delta_{in} = \sqrt{n}(\hat{F}_i - F_i). \end{aligned} \tag{B.4}$$

Notice that terms like  $\psi^\epsilon$ ,  $\lambda_i^\psi$ ,  $\lambda_i^\epsilon$  can be passed through the integral. Therefore when we evaluate at  $\epsilon = 0$ , we obtain

$$\psi^\epsilon(0) \sum_i E(U_{1i1}^\psi) + \sum_i \frac{\sum_j U_{1ij}}{\sqrt{n}} = 0$$

because  $E(U_{1i1}^{\lambda_i}) = 0$ . Therefore

$$\sqrt{p} \psi^\epsilon(0) = - \frac{\frac{1}{\sqrt{pn}} \sum_i \sum_j U_{1ij}}{\frac{1}{p} \sum_i E(U_{1i1}^\psi)} \xrightarrow{d} N(0, \bar{I}^{-1}).$$

The first order derivatives of  $U_1(\epsilon)$  and  $\tilde{U}_2(\epsilon)$  are the same at  $\epsilon = 0$ , because the first order derivative of the correction term in  $\tilde{U}_{2i}(\epsilon)$  always contains  $\epsilon$ , and hence equals zero at  $\epsilon = 0$ .

We will use the result  $\psi^\epsilon(0) = O_p(p^{-1/2})$  in the next section.

## B.4 Second order derivatives

We compute the second order derivative of  $U_1(\epsilon)$  first. Taking the derivative with respect to  $\epsilon$  on the first derivative given in equation (B.4), we have

$$\sum_i \left\{ \psi^{\epsilon\epsilon} \int \left( U_{1i1}^\psi + U_{1i1}^{\lambda_i} \lambda_i^\psi \right) dF_i(\epsilon) \right. \quad (\text{B}\cdot\text{5})$$

$$+ \psi^\epsilon \int \left( U_{1i1}^{\psi\psi} \psi^\epsilon + U_{1i1}^{\psi\lambda_i} \lambda_i^\psi \psi^\epsilon + (U_{1i1}^{\lambda_i\psi} \psi^\epsilon + U_{1i1}^{\lambda_i\lambda_i} \lambda_i^\psi \psi^\epsilon) \lambda_i^\psi \right. \\ \left. + U_{1i1}^{\lambda_i} \lambda_i^{\psi\psi} \psi^\epsilon \right) dF_i(\epsilon) \quad (\text{B}\cdot\text{6})$$

$$+ \psi^\epsilon \int \left( U_{1i1}^{\psi\lambda_i} \lambda_i^\epsilon + U_{1i1}^{\lambda_i\lambda_i} \lambda_i^\psi \lambda_i^\epsilon + U_{1i1}^{\lambda_i} \lambda_i^{\psi\epsilon} \right) dF_i(\epsilon) \quad (\text{B}\cdot\text{7})$$

$$+ \psi^\epsilon \int \left( U_{1i1}^\psi + U_{1i1}^{\lambda_i} \lambda_i^\psi \right) d\Delta_{in} \quad (\text{B}\cdot\text{8})$$

$$+ \int \left( (U_{1i1}^{\lambda_i\psi} \psi^\epsilon + U_{1i1}^{\lambda_i\lambda_i} \lambda_i^\psi \psi^\epsilon) \lambda_i^\epsilon + U_{1i1}^{\lambda_i} \lambda_i^{\epsilon\psi} \psi^\epsilon \right) dF_i(\epsilon) \quad (\text{B}\cdot\text{9})$$

$$+ \int \left( U_{1i1}^{\lambda_i\lambda_i} \lambda_i^\epsilon \lambda_i^\epsilon + U_{1i1}^{\lambda_i} \lambda_i^{\epsilon\epsilon} \right) dF_i(\epsilon) \quad (\text{B}\cdot\text{10})$$

$$+ \int U_{1i1}^{\lambda_i} \lambda_i^\epsilon d\Delta_{in} \quad (\text{B}\cdot\text{11})$$

$$+ \int \left( U_{1i1}^\psi \psi^\epsilon + U_{1i1}^{\lambda_i} \lambda_i^\psi \psi^\epsilon \right) d\Delta_i \quad (\text{B}\cdot\text{12})$$

$$+ \left. \int U_{1i1}^{\lambda_i} \lambda_i^\epsilon d\Delta_{in} \right\} \quad (\text{B}\cdot\text{13})$$

$$= 0.$$

Based on the regularity conditions 3 and 5, it can be seen that when evaluated at  $\epsilon = 0$ , only four terms are of order  $O_p(p)$ : the first term in (B.5), the first term in (B.10), and the identical terms in (B.11) and (B.13). The other terms are of order  $O_p(1)$ . The above equation can therefore be written in a simpler form:

$$\psi^{\epsilon\epsilon}(0) \sum_i E(U_{1i1}^\psi) + \left[ \sum_i \lambda_i^\epsilon \lambda_i^\epsilon E(U_{1i1}^{\lambda_i\lambda_i}) + 2 \sum_i \int U_{1i1}^{\lambda_i} \lambda_i^\epsilon d\Delta_i \right] + O_p(1) = 0.$$

The sum of the two terms in the brackets can be expressed as

$$2 \sum_i \left[ \left( \sum_j \frac{1}{\sqrt{n}} \frac{V_{1ij}}{E(V_{1i1}^2)} \right) \left\{ \sum_j \frac{1}{\sqrt{n}} \left( U_{1ij}^{\lambda_i} - \frac{E(U_{1i1}^{\lambda_i} V_{1i1})}{2E(V_{1i1}^2)} V_{1ij} \right) \right\} \right].$$

The terms in parentheses have joint limiting normal distributions by application of the multivariate central limit theorem. Their product is a quadratic form in normal random variables with mean  $-E(U_{1i1}V_{2i1})/E(V_{1i1}^2)$ . Therefore, for  $U_1(\epsilon)$ , we have

$$\frac{1}{2}\psi_1^{\epsilon\epsilon}(0) \xrightarrow{p} -\frac{1}{2} \lim_{p \rightarrow \infty} \sum_{i=1}^p \frac{E(U_{1k_j}V_{2k_j})/E(V_{1k_j}^2)}{p\bar{I}_p} = b.$$

Now consider the second order derivative of  $\tilde{U}_2(\epsilon)$ . For each stratum  $i$ , it is

$$-2\tilde{\rho}_i E(V_{1i1}^{\lambda_k}) = \frac{2E(U_{1i1}V_{2i1})}{2\{E(V_{1i1}^2)\}^2} E(V_{1i1}^2) = \frac{E(U_{1i1}V_{2i1})}{E(V_{1i1}^2)}.$$

Thus, for  $U_2(\epsilon)$  we have  $\frac{1}{2}\psi_2^{\epsilon\epsilon}(0) \xrightarrow{p} 0$  because the correction term exactly cancels the bias.

## B.5 Remainder term

### B.5.1 For expansion of $\hat{\psi}_1(\epsilon)$

We will evaluate the remainder term of the expansion. Assuming third order differentiability of  $\psi_1(\epsilon)$ , a Taylor expansion implies that

$$\sup_{0 \leq \epsilon \leq \frac{1}{\sqrt{n}}} \sqrt{pn} \left( \frac{1}{\sqrt{n}} \right)^3 |\psi_1'''(\epsilon)| \xrightarrow{p} 0.$$

Since  $n^{-1}p^{1/2} \rightarrow 0$ , we only need prove that

$$\sup_{0 \leq \epsilon \leq n^{-1/2}} |\psi_1'''(\epsilon)| = O_p(1).$$

Taking  $\epsilon$ -derivatives of (B.5)—(B.13), the coefficient of  $\psi_1'''$  in the expansion is

$$\sum_{i=1}^p \left\{ \int (U_{1ij}^\psi + U_{1ij}^{\lambda_i} \lambda_i^\psi) dF_{\epsilon_i} \right\}.$$

Thus the 3rd order derivative on  $U_1(\epsilon)$  can be written as

$$\psi_1''' \sum_{i=1}^p \left\{ \int (U_{1ij}^\psi + U_{1ij}^{\lambda_i} \lambda_i^\psi) dF_{\epsilon_i} \right\} + \sum_{i=1}^p \{\text{all other terms}\} = 0.$$

where each term in “all other terms” is at most  $O_p(1)$ . Therefore, we have

$$\sup_{0 \leq \epsilon \leq n^{-1/2}} |\psi'''(\epsilon)| = \sup_{0 \leq \epsilon \leq n^{-1/2}} \left| \frac{\sum_{i=1}^p \{\text{all other terms}\}}{\sum_{i=1}^p \left\{ \int (U_{1ij}^\psi + U_{1ij}^{\lambda_i} \lambda_i^\psi)(\psi, \epsilon) dF_\epsilon \right\}} \right|.$$

Since the numerator is at most  $O_p(p)$ , we only need to show that the denominator divided by  $p$  converges to a non-zero constant as  $p \rightarrow \infty$ . Specifically, for the denominator we will show that

$$\frac{1}{p} \sum_{i=1}^p \left\{ \int U_{1ij}^\psi(\psi, \epsilon) dF_\epsilon \right\} \xrightarrow{p} -\bar{I} \text{ uniformly for all } \epsilon \in [0, n^{-1/2}],$$

and

$$\frac{1}{p} \sum_{i=1}^p \left\{ \int U_{1ij}^{\lambda_i} \lambda_i^\psi(\psi, \epsilon) dF_\epsilon \right\} \xrightarrow{p} 0 \text{ uniformly for all } \epsilon \in [0, n^{-1/2}].$$

Since  $\bar{I}$  is assumed nonzero, we complete the proof.

The proof to the above two statements is essentially the same so we only prove the first statement. For each stratum  $i$ ,

$$\begin{aligned} & \int U_{1i1}^\psi(\psi(\epsilon), \lambda_i(\psi(\epsilon), \epsilon)) dF_\epsilon - \int U_{1i1}^\psi(\psi, \lambda_i(\psi, 0)) dF_0(x) \\ &= \left[ \int U_{1i1}^\psi(\psi(\epsilon), \lambda_i(\psi(\epsilon), \epsilon)) dF_\epsilon - \int U_{1i1}^\psi(\psi, \lambda_i) dF_\epsilon(x) \right] \\ & \quad + \left[ \int U_{1i1}^\psi(\psi, \lambda_i) dF_\epsilon - \int U_{1i1}^\psi(\psi, \lambda_i) dF_0(x) \right] \\ &= \int U_{1i1}^{\psi\psi}(\psi^*, \lambda_i^*) \cdot (\psi(\epsilon) - \psi) dF_\epsilon(x) \\ & \quad + \int U_{1i1}^{\psi\lambda_i}(\psi^*, \lambda_i^*) \cdot (\lambda_i(\psi(\epsilon), \epsilon) - \lambda_i) dF_\epsilon(x) \\ & \quad + \int U_{1i1}^\psi(\psi, \lambda_i) d[\epsilon\sqrt{n}(\hat{F} - F)], \end{aligned}$$

where  $\psi^*$  is an intermediate value between  $\psi(0)$  and  $\psi(\epsilon)$ , and  $\lambda_i^*$  is an intermediate value between  $\lambda_i$  and  $\lambda_i(\psi(\epsilon), \epsilon)$ .

If we add this expansion over strata and divide by  $p$ , the first term on the right hand side becomes

$$\frac{1}{p} \sum_{i=1}^p \int U_{1i1}^{\psi\psi}(\psi^*, \lambda_i^*) \cdot [\psi(\epsilon) - \psi] dF_\epsilon(x),$$

which is bounded by (using Assumptions 6, 7, 8, 10)

$$\begin{aligned} & \left| \psi(\epsilon) - \psi \right| \cdot \frac{1}{p} \sum_{i=1}^p \left| (1 - \epsilon\sqrt{n})E(M_i) + (\epsilon\sqrt{n})\frac{1}{n} \sum_{j=1}^n M_i(x_{ij}) \right| \\ & \leq \left| \psi(\epsilon) - \psi \right| \cdot \frac{1}{p} \sum_{i=1}^p \left[ E(M_i) + \frac{1}{n} \sum_j M_i(x_{ij}) \right] \xrightarrow{p} 0 \text{ uniformly in } \epsilon. \end{aligned}$$

The second term on the right hand side becomes

$$\begin{aligned} & \frac{1}{p} \sum_{i=1}^p [\lambda_i(\psi(\epsilon), \epsilon) - \lambda_i] \cdot \int U_{1i1}^{\psi\lambda_i}(\psi^*, \lambda_i^*) dF_\epsilon(x) \\ & \leq \left[ \frac{1}{p} \sum_{i=1}^p [(1 - \sqrt{n}\epsilon)E(M_i) + \sqrt{n}\epsilon\hat{E}_n(M_i)]^2 \right]^{\frac{1}{2}} \left[ \frac{1}{p} \sum_{i=1}^p [\lambda_i(\psi(\epsilon), \epsilon) - \lambda_i]^2 \right]^{\frac{1}{2}}. \end{aligned}$$

The first factor on the right converges in probability to  $\lim_{p \rightarrow \infty} \frac{1}{p} \sum_{i=1}^p [E(M_i)]^2$  by the law of large numbers for triangular arrays. The second factor converges to zero in probability by the conditions on the sequence  $\{\lambda_i\}$  given in Assumption 9. Therefore the second term goes to zero in probability. The third term goes to zero by the law of large numbers.

Therefore,

$$\frac{1}{p} \sum_{i=1}^p \left\{ \int U_{1ij}^{\psi}(\psi, \epsilon) dF_\epsilon \right\} \xrightarrow{p} -\bar{I} \text{ uniformly for all } \epsilon \in [0, n^{-1/2}].$$

### B.5.2 For expansion of $\hat{\psi}_2(\epsilon)$

The second order derivative of  $U_2(\epsilon)$  has one more term than that of  $U_1(\epsilon)$ , which comes from the correction term of  $U_2(\epsilon)$ . At  $\epsilon \in [0, n^{-1/2}]$ , it is

$$-2 \sum_{i=1}^p \tilde{\rho}_i(\psi(\epsilon), \lambda_i(\psi(\epsilon), \epsilon)) \int V_{1i1}^{\lambda_i}[\psi(\epsilon), \lambda_i(\psi(\epsilon), \epsilon)] dF_i(\epsilon) \quad (\text{B}\cdot 14)$$

Therefore, the third order derivative of  $U_2(\epsilon)$  has one more term than that of  $U_1(\epsilon)$ , which is the derivative of (B·14).

This term does not contain a factor of  $\psi'''$ , and is of order at most  $O_p(p)$ , so it can be included in the “all other terms” in above section. Therefore, the only essential difference

from the  $\hat{\psi}_1$  case is that here the estimator  $\psi(\epsilon)$  is actually  $\psi_2(\epsilon)$ . Since we have assumed the consistency of  $\psi_2(\epsilon)$  as well, we can use the same approach as for the  $\hat{\psi}_1$  case to show that the remainder term of  $\psi_2(\epsilon)$  is as small as that of  $U_1(\epsilon)$ .

## References

- Andersen, E. B. (1970a). Asymptotic properties of conditional maximum-likelihood estimators (corr: 71v33 p167). *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.
- Barndorff-Nielsen, O. E. (1980). Conditional resolutions. *Biometrika*, 67(2):293–310.
- Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70(2):343–365.
- Berger, J. O., Liseo, B., and Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters (with discussion). *Statist. Sci.*, 14(1):1–28.
- Boos, D. D. and Brownie, C. (1995). Anova and rank tests when the number of treatments is large. *Statist. Probab. Letters*, 23:183–191.
- Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, New York.
- Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63(2):277–284.
- Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.*, 5:815–841.
- Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion). *J. R. Statist. Soc. B*, 32:175–208.
- Lindsay, B. G. (1982). Conditional score functions: some optimality results. *Biometrika*, 69(3):503–512.
- Lindsay, B. G. (1995). *Mixture models: Theory, geometry and applications*. Institute of Mathematical Statistics, Hayward, California.
- Lindsay, B. G. and Waterman, R. (1999). Second-order information loss due to nuisance parameters: A simple measure. In Ghosh, S., editor, *Asymptotics, Nonparametrics, and Times Series*, pages 789–810, New York. Marcel-Dekker, Inc.

- Mak, T. K. (1982). Estimation in the presence of incidental parameters. *Canad. J. Statist.*, 10:121–132.
- McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *J. R. Statist. Soc. B*, 52(2):325–344.
- Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):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.
- Reid, N. (1996). Likelihood and Bayesian approximation methods. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics*, volume 5, pages 351–368. Oxford University Press.
- Small, C. G. and McLeish, D. L. (1988). Generalizations of ancillarity, completeness and sufficiency in an inference function space. *Ann. Statist.*, 16(2):534–551.
- Small, C. G. and McLeish, D. L. (1989). Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika*, 76(4):693–703.
- Small, C. G. and McLeish, D. L. (1994). *Hilbert Space Methods in Probability and Statistical Inference*. John Wiley, New York.
- Waterman, R. P. (1993). *Projected Score Methods*. PhD thesis, Pennsylvania State University.
- Waterman, R. P. and Lindsay, B. G. (1996a). Projected score methods for approximating conditional scores. *Biometrika*, 83(1):1–13.
- Waterman, R. P. and Lindsay, B. G. (1996b). A simple and accurate method for approximate conditional inference applied to exponential family models. *J. R. Statist. Soc. B*, 58(1):177–188.